Accuracy of interviewer observations: evaluating between-interviewer agreement in the National Health Interview Survey

Nancy Bates U.S. Census Bureau[1]

Rachael Walsh, U.S. Census Bureau

James Dahlhamer, National Center for Health Statistics

International Workshop on Household Survey Nonresponse

London, England September 4-6th 2013

August 19, 2013

[1] This report represents a work in progress and is released for purposes of discussion. The opinions expressed here are those of the authors and not necessarily the U.S. Census Bureau or National Center for Health Statistics.

Background

Interviewer observations collected during computer-assissted personal interviews represent a new type of paradata being collected in some surveys. These may include observations of sample unit neighborhoods, physical attributes of the units, or assessments of sample unit residents. These auxiliary data are being explored to inform adaptive designs and as covariates for non-response weighting adjustments (Kreuter, et al. 2010; Little and Vartivarian, 2005). Interviewer observations also provide contextual information about interviewer assignments and caseloads and may provide yet another input for response propensity models.

To inform adaptive designs, interviewers record observations on the first visit attempt – and in many cases before contact is made. As such, observations may involve educated guesses with discretion for how a judgment is made (e.g. assessing whether the unit contains young children). Because the systematic collection of observations is a new task for interviewers, more recent studies have begun to explore the error properties and factors influencing the accuracy of interviewer observations (West, 2013; Sinibaldi, Durrant and Kreuter, 2013; Casas-Cordero, Kreuter, Wang and Babey, 2013; West and Kreuter, 2013).

This paper examines observations recently collected in the U.S. National Health Interview Survey (NHIS) – a cross-sectional nationally representative personal visit survey sponsored by the National Center for Health Statistics with data collected by the U.S. Census Bureau. Beginning in January 2013, the NHIS added fifteen observations collected via laptops during the first personal visit contact attempt. For example, observations of neighborhood graffiti, physical condition of the sample unit, evidence of smokers, and evidence of household members with a chronic disability or health condition. In our analyses, we examine levels of between-interviewer agreement and explore the measurement error properties of various observations. To do this, we analyze a nonrandom subset of cases where multiple observations were collected for the same sample unit. We address the following research questions:

- What is the level of reliability between interviewers collecting observations for the same sample units and does it vary by the type of observation being measured?
- Does reliability vary according to the characteristics of the attempt when the observation is recorded, for example:
    - Whether same or different interviewers recorded the set of observations,
    - Whether access barriers were present, or
    - Whether contact was made?
- Do interviewer characteristics influence reliability?

The goals of collecting NHIS observations were twofold: (1) to serve as possible inputs for nonresponse weighting adjustments and (2) to serve as paradata inputs for various aspects of adaptive survey design, including propensity models. An overarching goal was to select a set of observations that might work across multiple demographic surveys fielded by the Census Bureau.

Tables 1 contains the 15 observations collected in the NHIS (see Miller, et al., 2013 for rationale and selection criteria).

Table 1 - Interviewer Observations tested in 2013 NHIS (in order they appeared)

- SCREENER: On this contact attempt, were you able to observe the sample unit or building within which the sample unit resides?
  1. Yes (continue with observations)
  2. No (skip observations)

- Did you observe graffiti or painted over graffiti on buildings, sidewalks, walls or signs in the block face or the sample unit or building within which the sample unit resides?

- How would you describe the condition of the sample unit our building within which the sample unit resides?

  1. Very Poor 2. Poor 3. Fair 4. Good 5. Excellent

- Based on your observation, does the sample unit or building within which the sample resides have:
  …a security buzzer/keycode/doorman or other barrier that may prevent access (e.g. dogs, locked gate)?
  …well-tended yard or garden?
  …peeling paint or damaged exterior walls?
  …window bars or grating on the doors or windows?
  1. Yes
  2. No

- Based on your observation, does the <u>SAMPLE UNIT</u> have:
  …3 or more door locks?
  …indication that children under 6 (including babies) may live at the unit (visible toys, car seat, strollers, outdoor swing/play set for example)?
  …a wheel chair ramp or other indicators that the residents of the sample unit are handicapped, disabled, or may have a chronic health condition (deaf, blind, use oxygen, etc.)?
  …an adult size bicycle?
  …any indication that the residents of the sample unit are smokers (cigarette/cigar butts, ashtrays, smell smoke?)
  1. Yes
  2. No
  3. Unable to observe sample unit (*option 3 presented only if buzzer/keycode/doorman=yes*)

- Relative to the general population and based on your observations, would you judge this sample unit to have a household income in the bottom third, middle third or top third of the population?

- Based on your observation, would you say at least one adult resident of the sample unit is employed?

- Based on your observation, would you say the residents of the sample unit speak a language other than English?

- How old would you estimate the residents of the sample unit to be?
  1. All occupants under the age of 30
  2. All occupants over the age of 65
  3. Other age composition

A screener item began the series and read: "*On this contact attempt, were you able to observe the sample unit or building within which the sample unit resides?*" If the answer is "no" interviewers skipped the observations during that contact attempt. For the five observations based specifically on the sample unit (LOCKS, CHILD, WHEELCHAIR, BIKE, SMOKE), in addition to yes/no, a third category of "unable to observe sample unit" was presented in situations where a security buzzer/keycode/doorman or other barrier was earlier noted. This allowed interviewers an "out" in cases where the access barrier prevented them from observing the actual unit.

Observations were programmed in a stand-alone Blaise instrument that was invoked when interviewers recorded contact histories for the first personal visit attempt. The instrument launched when the first personal visit attempt concluded, regardless of whether an interview occurred. Once observations were recorded, the items did not come on path again. Observations were recorded first, followed by the usual contact history items such as whether contact was made, outcome of the attempt, respondent concerns, and strategies used.

Limitations

First, unlike some interviewer observation studies, (most notably the Los Angeles Family and Neighborhood Study or LA-FANS), the NHIS interviewers had very limited training on how to collect the observations. Interviewers received a self-study training module that included a brief overview of new topics to the survey, including the observation measures. They also received a new observation training module at refresher training that consisted of a 60 minute interactive presentation.

Second, observations were intended to be collected only once for each sample unit. However, duplicate observations could occur under two conditions: (1) a case was reassigned to a different interviewer for purposes of refusal conversion follow-up, scheduling difficulties, vacation, or (2) a case had to be restarted by the same interviewer due to laptop malfunctions or other technical problems that resulted in a second set of observations by the same interviewer. Consequently, the subset of cases and interviewers available for analysis are not random and represent only a small percent (around 12 percent) of the NHIS production sample. We are mindful of this when drawing general conclusions.

Finally, because the NHIS survey data are not yet available, we cannot use the survey answers to determine accuracy of the observations. When these data become available, seven of the observations can be compared to questionnaire items to gauge metrics such as false positives and false negatives. For this paper, the reliability of observations is limited to traditional inter-observer reliability metrics such as between-interviewer agreement rates and kappa statistics. Differences noted throughout the paper (e.g., differences in observation agreement rates) reflect statistically significant differences at the .01 level.

Data

Table 1 – Final outcome of cases with duplicate observations versus all cases

	All cases with multiple observations	All Cases
Completed interview	35.5%	50.7%
Refusal	27.7	11.9
No one home	7.6	3.6
Other non-interview eligible	12.0	5.6
Ineligible/out of scope	17.3	28.3
	100%	100%
N	4,697	36,470

Source: 2013 NHIS Interviewer observations January – May

As expected, the outcomes for sample units with duplicate observations look different from all NHIS outcomes. Cases with duplicate observations were less likely to be interviews and more likely to be refusals and other types of eligible non-interviews (Table 1). Table 2 indicates that most duplicates were recorded by two different interviewers, likely due to reassignments for refusal conversion purposes. Overall, the majority of observations were recorded as instructed – on the first personal visit. Very few were recorded after 7 pm when it was likely dark and most were recorded immediately after the contact attempt. The average time spent collecting observations was 75 seconds with a mode of 34 seconds. Interviewers recorded the presence of some type of access barrier just under 20 percent of the time and reported making contact with an eligible household member on the observation attempt just over one-third of the time.

Table 2 - Descriptive Statistics of Duplicate Observations

Variable	Percent or mean
Occurrence of unique interviewer pairs	3.4 (mean)
	3 (mode)
Different interviewers	90.1
Reassigned	92.9
Recorded on 1st personal visit	80.3
Recorded in evening*	10.6
Recorded at time of attempt +	92.0
Time spent collecting (seconds)	74.5 (mean)
	34 (mode)
Presence of access barrier	18.1
Contact made during observation	35.6
N	4,071

Source: 2013 NHIS interviewer observations January-May. Cases with exactly 2 sets of observations.

**Evening defined as after 7 pm*

+Interviewers were allowed to record contact histories and observations after the contact attempt, for example, later in the evening at home.

Data in Table 3 suggest that the interviewers recording the second observation were more likely to be a supervisor and male and less likely to be an intermittent employee. Other characteristics appear similar between interviewer pairs.

Table 3. Descriptive Statistics of NHIS Interviewers

	Percent or mean INTERVIEWER 1	Percent or mean INTERVIEWER 2
Female	72.0%	63.7%
Supervisor	6.3	14.4
Intermittent Employee	53.1	43.0
Education		
High School	55.5	57.2
Some college	16.2	15.3
BA or higher	28.2	27.5
Years Census experience	4.3	4.7
N		

Source: Center for Adaptive Design employee dataset

Findings

Table 4 provides frequencies of reporting the presence of the various observations. Items are grouped according to a factor analysis (not shown). This analysis suggested four underlying factors including Appearance, At home-ness, Security, Activity/Children and a fifth category for two items that did not load on other factors. For most items within a factor, the percent presence of an observation is presented, with the exception of means for address condition scores (on a 5 point scale) and household income average (3 point scale). Some items were rarely recorded as being present, e.g., evidence of health problems (2.7 percent), an adult bicycle (2.5 percent), smokers (5.7 percent) and graffiti (4.5 percent). These may suggest propensity for false negative rates. Alternatively, presence of an employed household resident was rather high at 80 percent.

Table 4. Percentage distribution of observations (unweighted)

Items	Label	Type[a]	% or mean
Appearance items			
1 Address condition (5 pt. scale)	COND	O	3.8
2 Well-kept yards	YARD	D	63.2
3 Damaged walls	WALL	D	12.3
4 Income (3 ordered categories)	INC	O	1.8
At home-ness items			
5 Wheelchair/health problems	WHEEL	D	2.7
6 Employed adult	EMP	D	80.0
7 HHD aged >65	OLDER	D	10.8
Security items			
8 Bars on windows	BARS	D	6.9
9 Multiple locks	LOCKS	D	4.4
10 Access barrier to sample unit	BARR	D	18.1

Active/Children items			
11 Adult bicycle	BIKE	D	2.5
12 Children <6	CHILD	D	11.6
Misc. items			
13 Evidence of smokers	SMOKE	D	5.7
14 Graffiti on block face	GRAF	D	4.5
15 Evidence another language spoken	LANG	D	16.2

[a] Type of item: D=dichotomous; O=ordinal .
Source: 2013 NHIS interviewer observations January-May. Cases with exactly two sets of observations.

Figure 1 illustrates the percent agreement across observations by same versus different interviewer pairs.[2] Overall, the absolute agreement rates are higher when the same interviewer recorded the second observation, but for most items, agreement levels are not significantly different regardless of whether the pair was same or different.[3] Notable exceptions are three items in the Appearance group – namely Address Condition, Well-Kept Yards, and Income bracket. Two of these (COND and INC) likely have greater variance because they are more subjective and use ordered categories as opposed to simple yes/no . The final item, YARD, contained a third category of "not applicable" for situations such as urban highrises lacking a common green area. Upon investigation, we found that forty-three percent of the disagreements for YARDS involved discrepancies around the "not applicable" category. We view the lack of differences as a positive finding and for the remainder of the analysis, concentrate on the 90% of cases with *different* interviewer pairs.

[2] For 5 items (LOCKS, BIKE, CHILD, SMOKER, WHEEL) a third category of "could not observe sample unit" appeared if the interviewer indicated that an access barrier was present. In these situations, interviewers could indicate presence or absence of the item or select the "could not observe" category. When computing agreement rates for Fig 1, we subset these items to situations where neither interviewer indicated presence of an access barrier.
[3] Statements indicating differences in agreement rates for Fig.1-3 reflects significant differences at the .01 level.

Figure 1. Percent observation agreement by same versus different interviewers

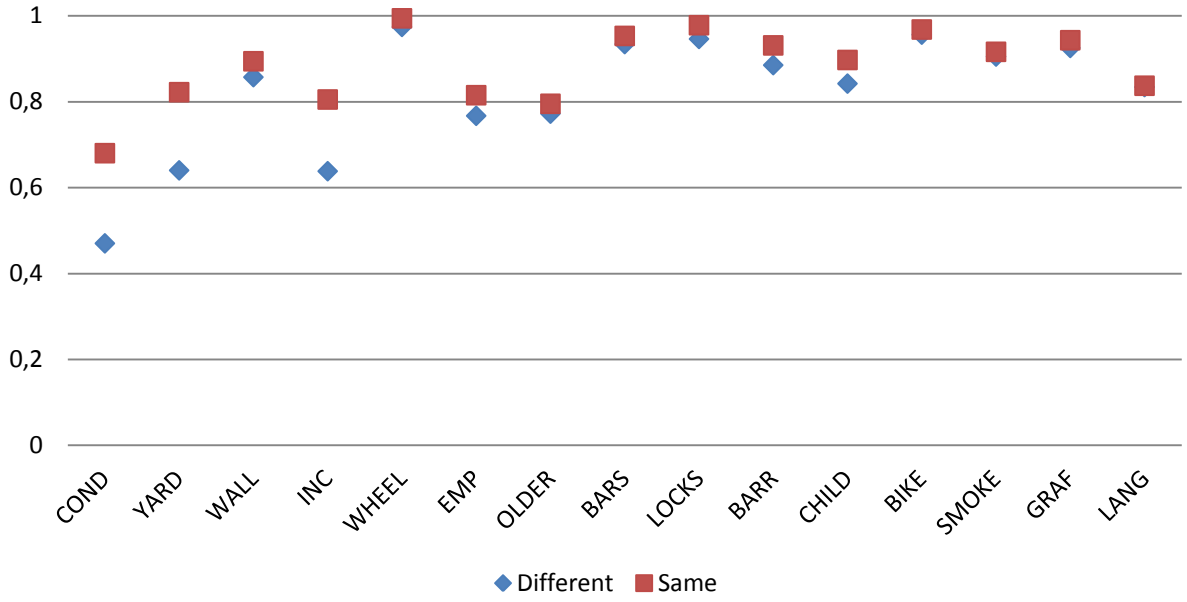


Figure 2 illustrates agreement rates given the presence or absence of an access barrier. We define an access barrier to be present if *both* interviewers indicated so. In situations where an access barrier was reported, the instrument was programmed to present a third response option for a subset of observations. The option "Unable to observe the sample unit" applied to WHEEL, LOCKS, CHILD, BIKE, and SMOKE. In these situations, interviewers could still select "yes or no" in cases where a barrier was noted, but did not prevent access to the sample unit.

The agreement rate across interviewers was conditional upon the presence or absence of an access barrier for some items. Three of the Appearance items (COND, WALL, INC) were not significantly different – this makes sense given most can be assessed regardless of security fences, buzzer entries and the like. However, significant differences in agreement rates were found for WHEEL, LOCKS, CHILD, BIKE and SMOKE – the subset of items where the "unable to observe" category was offered. A closer examination of these items revealed that, in the presence of an access barrier, most of the disagreement comes from situations when one interviewer answered yes or no, but the other answered "Unable to observe the sample unit". Therefore, most of the lower agreement rates noted in Figure 2 are not reflective of conflicting assessments, but rather situations where one interviewer provided an assessment while the other did not.

Fig. 2. Percent observation agreement by access barriers (different interviewers)

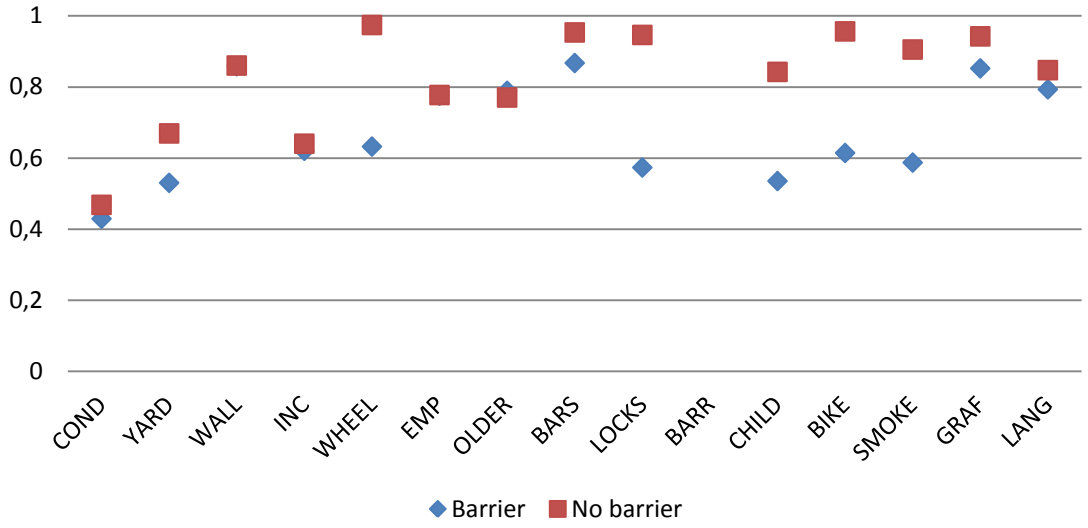


Figure 3 presents agreement rates by contact outcome status. We plot three outcomes – situations where both interviewers made contact, neither did, or one made contact while the other did not. Somewhat surprisingly, we see relatively few significant differences in agreement rates between outcomes. We do note, however, slightly (and significantly) lower agreement rates when one interviewer made contact but the other did not for: YARD, OLDER, LOCKS, CHILD, and LANG. This is consistent with findings reported by Walsh, Dahlhamer and Bates (2013) where answers to some observations appear to differ by contact versus no contact. Nonetheless, we view the overall lack of differences as favorable in terms of measurement error -- especially given that on our observation visits, contact was made just over one-third of the time. This suggests that interviewers can use other strategies such as neighborhood cues to make judgments. The lack of difference may also bode well for observer consistency across both responding and nonresponding households.

Fig. 3. Percent agreement by contact outcome
(subset to different interviewer pairs)

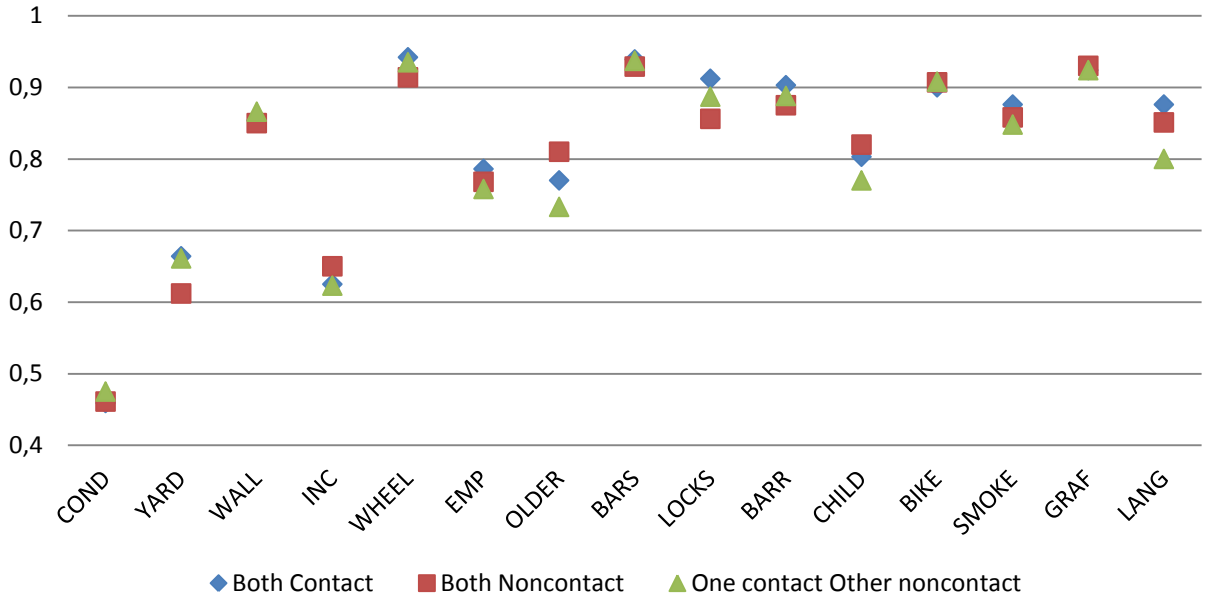
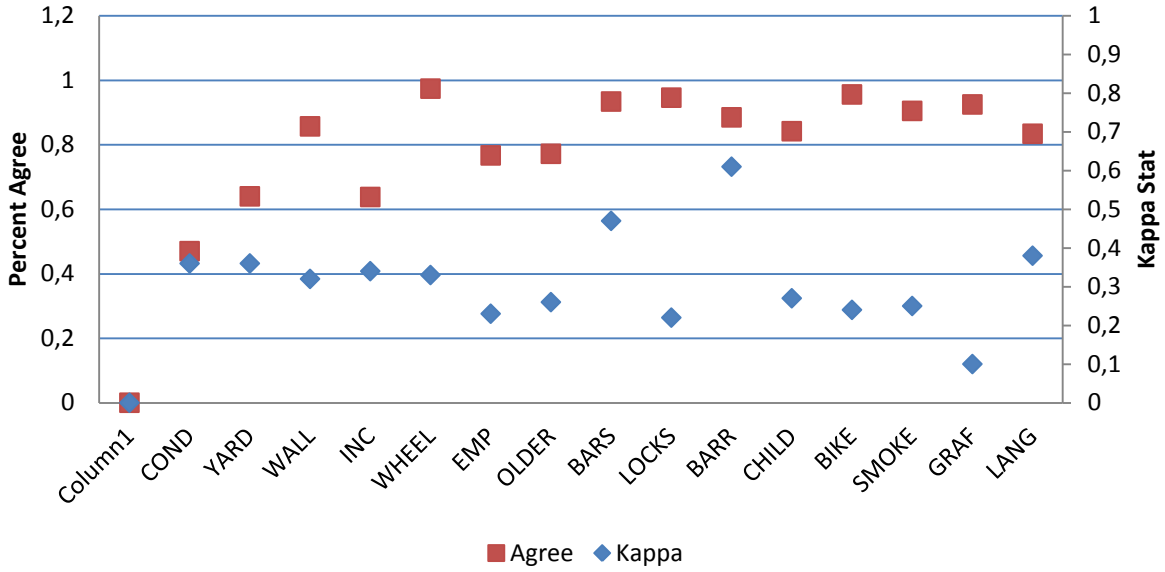


Figure 4. plots both the unadjusted agreement rates and kappa statistics across observations. In this case we report Cohen's Kappa, a statistic commonly used to indicate agreement rates for dichotomous data.[4] Kappa provides a quantitative measure of the magnitude of agreement between observers and takes into account the degree to which two observations might agree purely by chance (Viera and Garret, 2005). The figure clearly shows the kappa statistics are far below agreement rates, most in the .20-.35 range. The observation for GRAF is lowest at only .10 and it is interesting to note this item was the only neighborhood block-face level observation. This may suggest a need for better training definitions of a block face and/or graffiti.

Thresholds to determine an acceptable agreement rate using kappa vary, but most would place ours in the fair to moderate range. Both BARS and BARR are on the higher end so might be considered the most reliable (at least among those that cannot be validated by survey data).

[4] We report weighted kappa for COND since it was answer along a 5 point scale. All others are unweighted kappa.

Fig. 4. Interobserver agreement rates and kappa statistics
(subset to different interviewers pairs)



Our final table (table 5) presents results from a logistic regression model predicting agreement on a single observation – whether an adult in the household is employed. We ran a random effects model with only the intercept modeled randomly. We present 3 models. First is the unconditional model used as a reference for the second model which adds interviewer characteristics and third which adds situational variables of the observation. Interviewer characteristics include gender, education, years of experience with the Census Bureau, supervisory status, and whether a permanent or intermittent employee. Situational variables include whether the observation was recorded on the first personal visit, recorded immediately after the visit, the time spent recording the observations, whether contact was made on the observation visit, and whether an access barrier was present. To simplify analysis, we recoded most variables into dichotomous variables such that 1=both interviewers matched on a given characteristic/situation and 2=if different (see table 5).

Table 5 – Logistic regression model for inter-observer agreement for Employed

	Model 1	Model 2	Model 3
Fixed Effects	coefficient	Coefficient	coefficient
Interviewer characteristics			
Both Female		0.02	0.07
Education level same		-0.10	-0.04
Both years exp. above mean		-0.12	-0.21
Intermittent status same		0.01	0.03
Supervisory status same		0.05	0.10
Condition of Observation			
Both recorded on 1st PV			0.29**
Both immediately recorded			0.08
Both made contact			0.11
Neither made contact			0.15
One contact/other noncontact (omitted)			----
Both reported access barrier			-0.06
Time spent recording => mean			-0.30
Intercept	1.254***	1.279***	0.911***
Variance Components			
Interviewer pair	0.729*	0.734*	0.667*
Summary Statistics			
n	3201	3201	2841
Generalized chi-sq / df	0.81	0.81	0.81

Source: 2013 January – May NHIS interviewer observations, January – May NHIS Contact History Instrument data, and Office of Survey Analytics Census interviewer dataset.
*.05, **.01, ***.001

For the employment observation (EMP), none of the interviewer characteristics were statistically significant in predicting agreement between interviewers. Model 3 incorporates covariates associated with the observation itself and only one was found to be significant – odds of agreement were higher in situations where both interviewers collected observations on the first personal visit. We also note that the variance of the interviewer pairs is significant in all three models, indicating that some interviewer pairs did better than others in terms of agreement, even after controlling for interviewer and observation characteristics.

We ran similar models but substituted agreement on the presence of children (CHILD) as the dependent variable (data not shown)[5]. We found similar results for models 1 and 2, but found two of the observation condition covariates to be significant in model 3. Specifically, we found significantly higher odds of agreement between interviewers reporting presence of young children in situations where *neither* interviewer made contact (compared to situations where one made contact but the other did

[5] In order for the CHILD model to converge in PROC GLIMMIX, we had to subset the data to interviewer pairs that occurred at least two times in the dataset. This reduced the Model 3 N to 2075.

not). Significantly lower agreement was found in situations where both interviewers reported an access barrier (compared to cases when neither reported it).

Discussion

Our analysis adds some new data toward understanding measurement error of interviewer observations. The observations in our study range from neighborhood items to sample unit characteristics to characteristics of the household residents. Some are more subjective than others which raises the question of variance between interviewers and whether items can be measured with any degree of precision. Some of our observations can be validated with survey data, while others cannot. For now, the survey data are not available so we rely on proxy measures such as agreement rates and kappa statistics.

For most items, we found little difference in agreement rates regardless of whether the same or different interviewers recorded the duplicate observations. As might be expected, we saw higher disagreement rates when scale response options were used versus simple yes/no options. We also found relatively high agreement rates even when an access barrier was reported (and where we did find the observations did not match, it didn't usually reflect true disagreement). For the two observations we chose to model -- whether an adult member was employed and evidence of young children – we found little in the way of significant predictors to account for variance in agreement. The lack of interviewer-level features having significant effect on agreement is interesting when considering findings in West and Krueter (2013). In that study, the authors report that features of interviewers had no effect on the *accuracy* of observations. If we equate agreement with accuracy, then our findings seem in step with theirs. The general lack of predictive power for our situational observations also seems in line with at least one other study (Cordero, Kreuter, Wang and Babey,2012), where covariates such as rating after 5 pm and day of week were not found to predict observations for trash and graffiti.

Finally, three key pieces of information for our study are yet to be known. First, just how accurate are the observations we can validate with survey data? Second, do any of the observations correlate with key NHIS health indicators? Third, do any observations correlate with response propensity? Very preliminary data on the latter suggest that in the absence of other paradata predictors, most of the observations are significantly predictive in contact level response models (Erdman and Dahlhamer, 2013). Among the highest ranking are: YARDS, LANG, SMOKER, and INC. This is encouraging especially since SMOKER seems likely to also correlate with health outcomes. With the exception of SMOKER, our data indicate each of these observations have fair agreement rates, as measured by the kappa statistic.

The evidence from this analysis suggests that the observation instrument is generally working well in providing consistent information about housing units. There is some indication for modifications in wording or training that may improve reliability. In general, further evaluation is warranted for the use of these observations in adaptive design and nonresponsive adjustments.

References

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.

Erdman, C. and Dahlhamer, J. (2013). Evaluating Interviewer Observations in the National Health Interview Survey: Associations with Response Propensity. A paper presented at the Joint Statistical Meetings, ASA, Montreal.

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.

Little, R. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31:161-168.

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 . A paper presented at the Joint Statistical Meetings, ASA, Montreal.

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

Viera, A.J. and Garrett, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine, 37 (5): 360-3.

Walsh, R., Dahlhamer, J., Bates, N. (2013). Assessing Interviewer Observations in the NHIS. A paper presented at the Joint Statistical Meetings, ASA, Montreal.

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.