Published in final edited form as:

Methodology (Gott). 2018 January ; 14(1): 16–29. doi:10.1027/1614-2241/a000142.

Strategies for Increasing the Accuracy of Interviewer Observations of Respondent Features: Evidence from the U.S. National Survey of Family Growth

Brady T. West and

Survey Research Center, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48104

Frauke Kreuter

Joint Program in Survey Methodology, 1218 LeFrak Hall, College Park, MD, 20742 and University of Mannheim, Mannheim, Germany and Institute for Employment Research, Nuremberg, Germany

Summary

Because survey response rates are consistently declining worldwide, survey researchers strive to obtain as much auxiliary information on sampled units as possible. Surveys using in-person interviewing often request that interviewers collect observations on key features of all sampled units, given that interviewers are the eyes and ears of the survey organization. Unfortunately, these observations are prone to error, which decreases the effectiveness of nonresponse adjustments based on the observations. No studies have investigated the strategies being used by interviewers tasked with making these observations, or examined whether certain strategies improve observation accuracy. This study is the first to examine the associations of observational strategies used by survey interviewers with the accuracy of observations collected by those interviewers. A qualitative analysis followed by multilevel models of observation accuracy show that focusing on relevant correlates of the feature being observed and considering a diversity of cues are associated with increased observation accuracy.

Keywords

Interviewer Observations; Auxiliary Variables; Survey Paradata; Multilevel Modeling; Interviewer Effects

Interviewers in “face-to-face” surveys are often tasked with observing key features of sampled units, given that the interviewers are the eyes and ears of the survey organization. Because response rates in surveys of all formats have been consistently declining worldwide (Baruch and Holtom, 2008; Biener et al., 2004; Cull et al., 2005; Curtin et al., 2005; de Leeuw and de Heer, 2002; Tolonen et al., 2006; Brick and Williams, 2013; Williams and Brick, 2017), interviewers are asked to do this in an effort to obtain as much relevant auxiliary information on all sampled units as possible. The survey methodology literature has clearly established the need for auxiliary variables used for nonresponse adjustment of survey estimates to be related to both survey variables of interest and response propensity (Beaumont, 2005; Bethlehem, 2002; Groves, 2006; Lessler and Kalsbeek, 1992; Little and

Vartivarian, 2005; Kreuter et al., 2010), so that the adjustments will reduce both the bias and the variance in the ultimate survey estimates. Survey organizations will therefore request that interviewers attempt to collect observations on auxiliary variables having these optimal properties (in theory) for both respondents and nonrespondents, and then evaluate the potential of the observations for nonresponse adjustment purposes (e.g., West, 2013a).

Unfortunately, interviewer observations can be prone to error. The various conceptualizations of total survey error (TSE) that have been published over the years (Groves and Lyberg, 2010) consistently acknowledge the problem of nonresponse bias that can arise in surveys. Errors in estimation are less often acknowledged as a key part of TSE (Deming, 1944; Biemer, 2010). From a TSE perspective that also considers errors in estimation, errors in interviewer observations may lead to nonresponse adjustments that introduce *more* bias in survey estimates than was present before the nonresponse adjustments (Lessler and Kalsbeek, 1992; Stefanski and Carroll, 1985; West, 2013a; West, 2013b). This underscores the need for interviewer observations to be of sufficient accuracy, so that nonresponse adjustments based in part on the observations will do an effective job of repairing nonresponse bias. Researchers have started to consider design strategies for improving observation accuracy as a result (West and Kreuter, 2015).

More generally, the survey methodology literature has recently begun to examine the non-negligible error properties of these observations (Campanelli et al., 1997; Casas-Cordero et al., 2013; Groves et al., 2007; McCulloch et al., 2010; Pickering et al., 2003; Sinibaldi et al., 2013; Tipping and Sinibaldi, 2010; West, 2013a; West and Kreuter, 2013; West et al., 2014; West and Kreuter, 2015). Several recent studies have suggested that interviewers working on the same survey and receiving the same general training will vary substantially in terms of the accuracy of these observations, even when controlling for relevant respondent-, interviewer-, and area-level covariates (Sinibaldi et al., 2013; West and Kreuter, 2013; West et al., 2014; West and Kreuter, 2015).

So why might interviewers working on the same study, and having received the same training, vary substantially in terms of observation accuracy (even after adjustment for covariates that might influence the accuracy)? No previous studies have assessed the possibility that interviewers may be using different *strategies* to collect these observations (i.e., looking for different cues that might serve as indicators of a particular feature being observed), and considered whether different strategies in a given context may lead to more or less accurate observations. Because survey organizations may use these interviewer observations for nonresponse adjustment purposes, and interviewers have repeatedly been demonstrated to vary substantially in terms of the accuracy of the observations, this study sought to 1) explore the variability in the strategies that interviewers use to collect these observations, and 2) understand whether there is a relationship between the strategy used and observation accuracy. The identification of strategies that are associated with more accurate observations has clear implications for improving interviewer training, and the overall improvements in accuracy that may arise from more standardized training on this process may in turn improve the effectiveness of nonresponse adjustments based in part on these observations.

Why might interviewers vary in terms of the observational strategies that they use? In the absence of standardized training on the observation process, survey interviewers might utilize different “native” observational strategies, based on varying prior expectations of how the social world around them is organized (e.g., “a respondent who is currently in a romantic relationship will have features X, Y, and Z, so I will only look for features X, Y, and Z when making this judgment...”; Tversky and Kahneman, 1974; McCall, 1984; Funder, 1987; Manderson and Aaby, 1992a; Manderson and Aaby, 1992b; Funder, 1995). However, empirical evidence of this assumption is largely lacking in the methodological literature. Do interviewers really vary in terms of the cues that they are looking for when attempting to record these types of observations? From a psychological perspective, interviewer perceptions may be influenced by specific environmental and behavioral cues (Forgas and Brown, 1977). The literature cited above suggests that interviewers with different backgrounds and expectations may be influenced in different ways by different types of cues (both when recording observations and survey responses), meaning that only certain types of cues will influence the observations that they record. In the context of the present study, these more influential cues would define the “strategies” employed by interviewers when recording their observations. No studies to date have considered whether these naturally-varying strategies may ultimately lead to variability among interviewers in observation accuracy.

What observational strategies might be associated with increased accuracy? Some interviewers may resort to considering features of the areas / environments in which they are working in the absence of any household-specific cues. This could be helpful if the interviewers are familiar with the areas and the areas are fairly homogeneous in terms of the feature(s) being observed. But if areas tend to be more heterogeneous, interviewers may incorrectly apply expectations that all households in that area will have the same features, or assume that if several households have been similar, the next will also have similar features (Tversky and Kahneman, 1974; Seidler, 1974; Das, 1983; Manderson and Aaby, 1992a; Manderson and Aaby, 1992b; Harris et al., 1997; Repp et al., 1998; Millen, 2000; Babbie, 2001).

For other interviewers, the observational task may be quite difficult (e.g., inability to access a locked building of apartments, working in crowded urban areas, etc.), leading to a failure to pick up on important external visual cues and subsequent guessing or “going on hunches.” In these situations, observations would be expected to have *reduced* accuracy (Feldman et al., 1951; Kazdin, 1977; Jones et al., 1979; Graham, 1984; Funder, 1987; Most et al., 2005; Simons and Jensen, 2009). Other interviewers may attempt to pick up on several different relevant predictors of a feature being observed as well as specific features of a given household or respondent. These strategies reflecting diversity in the cues used (depending on the context) and an ability to detect specific features of a given respondent would be expected to have *increased* accuracy (Kazdin, 1977; Funder, 1995; West et al., 2014; West and Kreuter, 2015). The social psychology literature also suggests that observations based on first impressions in the presence of limited information will tend to have increased accuracy (Ambady et al., 1999; Patterson and Stockbridge, 1998). Whether or not these theoretical expectations related to accurate observational strategies are borne out in the survey interviewing context remains an open research question.

This article presents an examination of observational strategies that are associated with the accuracy of a key interviewer judgment in the U.S. National Survey of Family Growth (NSFG). NSFG interviewers (each of whom are female) first attempt to conduct a screening interview with an adult informant within a randomly sampled housing unit. The primary purpose of the screening interview is to determine whether the sampled housing unit contains a person between the ages of 15 and 49 (the target population for the NSFG). Upon identification of all eligible persons within a household, one age-eligible person is selected at random for the main (face-to-face) NSFG interview (possibly at a later date that is convenient for the selected respondent). Only about 77% of these selected individuals ultimately participate in the main interview, and this number has continued to decline gradually over time. Beginning with the seventh cycle of the NSFG (June 2006 – June 2010; Lepkowski et al., 2010), interviewers were tasked with judging whether this randomly selected age-eligible person was currently in a sexually active relationship with a member of the opposite sex, immediately after completion of the screening interview.

Why might the NSFG interviewers be asked to record this type of subjective judgment after each screening interview, given that this is not a directly observable feature of potential respondents? Unlike other types of auxiliary information commonly recorded by interviewers (e.g., contact observations, socio-demographics, housing unit features, etc.), this NSFG-specific judgment has a strong association with a number of key NSFG variables, along with the propensity of persons screened in the NSFG to respond to the main survey request (West, 2013a). These judgments therefore define an ideal auxiliary variable for nonresponse adjustment purposes (Little and Vartivarian, 2005). The NSFG in part focuses on several key variables related to sexual behavior and activity, and being able to get some indication from the interviewers as to whether *potential* respondents (selected from the initial screening interview) are currently in a sexually active relationship provides useful auxiliary information for persons who do not eventually respond to the main NSFG interview. NSFG staff can use this information (along with other auxiliary variables on the sampling frame) to predict key outcomes related to being in sexual relationships for main interview nonrespondents. These types of “tailored” interviewer observations, which survey managers can specifically design as potential correlates of key survey variables of interest, have also been shown to be more effective for nonresponse adjustments than other data sources, including linked information from commercial data sources (Sinibaldi et al., 2014).

These interviewer judgments could also be validated based on actual respondent reports of sexual activity collected in the main NSFG interview, which is important for assessing their accuracy. In addition to being asked to record these judgments after completing screening interviews, NSFG interviewers in the last two quarters of Cycle 7 data collection were also asked to provide open-ended justifications for why they made their judgments, and what cues they noted when recording the judgments. This study sought to leverage this qualitative information and assess the interviewer-specific observational strategies evident in these justifications, along with the amount of variability among interviewers in judgment accuracy that was explained by these strategies.

Given that no prior research has examined observational strategies being used by field interviewers tasked with making these types of judgments, the present study aimed to answer the following two research questions:

1. Do NSFG interviewers tend to fall into distinct “strategy” clusters based on the justifications used for their sexual activity judgments?
2. Were certain observational strategies associated with increased accuracy of the sexual activity judgments?

Data

Coding of Open-Ended Justifications

In the last two quarters of data collection for the NSFG (Cycle 7), 45 interviewers were asked to record (on laptop applications) open-ended justifications for their post-screener judgments of perceived current sexual activity for selected persons (see Figure 1). The interviewers were trained to provide the justifications immediately after the judgments were made, and the interviewers could not proceed with main interview tasks until after the judgments and their justifications had been entered (along with all other observations from the screening interview). This means that all judgments and justifications were recorded *prior* to the main NSFG interview, and there were no missing data for the judgments or justifications. The interviewers were not prompted for specific justifications or limited in any way (e.g., justification length), and interviewer training sessions on this process did not suggest any specific strategies to use when recording the judgments. The interviewers were simply told to make their best judgments based on what they had seen and/or heard, opening the door for the use of the aforementioned “native” observational strategies.

In total, the 45 interviewers provided 3,992 open-ended justifications of widely varying lengths during these two quarters of data collection. Two examples of real recorded justifications follow:

1. *“He works and goes to school and lives here with his twin - I do not think he could have someone over as the carpet is all taken up and it smells badly of dog poo.”* (A justification for a judgment of *not* currently sexually active, reflecting the use of cues describing features of the housing unit.)
2. *“He has a tattoo, ‘Carol’, over his heart.”* (A justification for a judgment of currently sexually active, reflecting an ability to pick up on the respondent’s appearance during the screening interview.)

The 3,992 justifications were coded on 13 different indicator variables (1 = mentioned in justification, 0 = not mentioned), with all indicators coded for each justification:

- Living arrangement (living with spouse, parents, etc.)
- Relationship status (mention of spouse, partner, etc.)
- Age
- Housing unit characteristics (presence of children, cleanliness, etc.)

- Appearance (references to physical appearance, ethnicity, etc.)
- Neighborhood characteristics
- Shyness
- Going on Hunches / Guessing (indication of a gut feeling, or not being sure)
- Incorrect (an incorrect judgment was entered in hindsight; after recording the judgment and having it saved in the system, the interviewer realized that they entered an incorrect judgment while they were writing their justification and mentioned this in their justification; we note that NSFG interviewers cannot go back and change their prior entries)
- Conservative (a conservative or strict household / parents)
- Health (reference to health or physical disability)
- Personality (reference to the person's personality)
- Occupation (reference to the person's occupation)

In addition, the number of words used for each justification was coded as a proxy of effort dedicated to the observational task. For example, the first justification given above was coded as having 35 words, and assigned a 1 for living arrangement, household characteristics, and occupation, and a 0 for all other indicators. All coding of the justifications was performed twice with the assistance of an undergraduate research assistant. The inter-rater reliability of the codes was quite high; of the 3,992 coded justifications, the percentages of codes on a given indicator that did not agree between the two coders ranged from 0.32% (Shyness) to 7.53% (Housing Unit Characteristics). This means that there was higher than 92% agreement for all of the coded indicators. Discrepancies in coding were detected using PROC COMPARE in the SAS software, and any discrepancies in coding or word counts were discussed and resolved. The percentage of justifications falling into each of these thirteen categories was then computed for each interviewer, along with the mean word count for the interviewer.

Measures of Observation Accuracy

To compute dependent variables measuring observation accuracy for each interviewer, the total number of judgments of sexual activity, the total number of sexually active respondents (based on respondent reports of at least one sexual partner in the past year, from completed main interviews), the total number of sexually inactive respondents, and the total number of *discordant* judgments (i.e., judgments inconsistent with survey reports) were determined for each of the 45 interviewers. From these measures, we computed the overall *gross difference rate* (i.e., the proportion of judgments that were incorrect), the *false positive rate* (i.e., the proportion of sexually inactive respondents who were judged to be sexually active), and the *false negative rate* (i.e., the proportion of sexually active respondents who were judged to not be sexually active) for each interviewer. For purposes of this study, respondent reports of current sexual activity (i.e., at least one opposite-sex partner in the past year) were assumed to be true.

Two (2) interviewers had insufficient information available for the false positive rate analyses, because all of their main interview respondents reported being sexually active. An additional two interviewers were found to be outliers in the cluster analyses and removed from further analysis. Our models of observation accuracy were thus based on 43 interviewers and 3,044 observations (for gross difference rates), 43 interviewers and 2,347 observations (for false negative rates), or 41 interviewers and 697 observations (for false positive rates).

Covariates

We also extracted a number of covariates describing features of the areas where an interviewer was assigned to work, motivated by prior studies examining correlates of observation accuracy (Sinibaldi et al., 2013; West and Kreuter, 2013; West and Kreuter, 2015). These included the percentage of the interviewer's assigned households in urban areas (i.e., Metropolitan Statistical Areas, or MSAs), the percentage of households in areas with access problems (e.g., gated communities), the percentage of households in primarily residential areas, the percentage of households in areas with evidence of non-English speakers or Spanish speakers, the percentage of households in areas with safety concerns, the percentage of households in multi-unit buildings, the percentage of households with physical impediments (e.g., security gates), and the percentage of households where females were the selected respondents. We describe the roles of these covariates in our models of accuracy below.

Analytic Approach

To address our first research question, we performed an exploratory cluster analysis to determine whether distinct groups of interviewers existed in terms of the percentages of justifications falling into each category and effort spent on the observational task. To do so, the 13 percentages and the mean word counts for the 45 interviewers who completed main interviews were initially standardized. An agglomerative hierarchical clustering approach was then applied (Everitt et al., 2011), using squared Euclidean distances based on the 14 standardized variables as distance measures between interviewers and Ward's (1963) minimum within-cluster variance method to define the clusters. This approach was selected for its established superiority in identifying known clusters when using continuous measures (Punj and Stewart, 1983). We examined descriptive statistics for the derived clusters, and then determined conceptual labels for the clusters by comparing the distributions of the percentages and means between them using the non-parametric, independent samples Kruskal-Wallis H test.

Next, to address our second research question, we fit a sequence of two multilevel logistic regression models to each of the three dependent accuracy measures (the gross difference rates, false positive rates, and false negative rates). The first model included random interviewer effects, capturing between-interviewer variability (and within-interviewer correlation) in a given accuracy indicator, and fixed effects of the $P = 9$ covariates described above, to account for the effects of these area-level features on the different accuracy measures. This initial model is shown in (1) below:

$$\ln \left(\frac{\phi_i}{1 - \phi_i} \right) = \beta_0 + \sum_{p=1}^P \beta_p x_{pi} + u_{0i} \quad (1)$$

where

ϕ_i = probability of incorrect judgment / false-positive / false-negative for IWER i

x_{pi} = value of covariate p for IWER i

$u_{0i} \sim N(0, \tau^2)$, with τ^2 = variance of random interviewer effects

Next, given an initial estimate of the variance of the random interviewer effects, we added fixed effects of the specific interviewer “strategy” clusters identified from the cluster analysis to the initial model (omitting one cluster as a reference category):

$$\ln \left(\frac{\phi_i}{1 - \phi_i} \right) = \beta_0 + \sum_{p=1}^P \beta_p x_{pi} + \sum_{c=1}^{C-1} \lambda_c I_i[\text{cluster} = c] + u_{0i} \quad (2)$$

where

ϕ_i = probability of incorrect judgment / false-positive / false-negative for IWER i

x_{pi} = value of covariate p for IWER i

$$I_i[\text{cluster} = c] = \begin{cases} 1 & \text{if IWER } i \text{ in cluster } c \\ 0 & \text{otherwise} \end{cases}$$

$u_{0i} \sim N(0, \tau^2)$, with τ^2 = variance of random interviewer effects

We then computed the percentage of variance in the random interviewer effects explained by the inclusion of the fixed effects of the strategy clusters in (2). In each model, the variance of the random interviewer effects was tested against zero using a likelihood ratio test based on a mixture of chi-square distributions (Zhang and Lin, 2008).

Models (1) and (2) were fitted using the GLIMMIX procedure in the SAS software (Version 9.4), specifically using a Laplace approximation for estimation purposes (Kim et al., 2013).

Results

We first consider descriptive summaries of the variables computed for all 45 interviewers. Descriptive statistics for the interviewer-specific percentages and mean word counts are shown in Table 1. A large amount of variability among the 45 interviewers is evident, in terms of the justification strategies and the average number of words used for the justifications.

Table 2 presents descriptive statistics for the three dependent variables computed for each interviewer. Interviewers made correct judgments of current sexual activity 79.4% of the time, with a minimum of 0% errors (i.e., some interviewers were correct all the time) and a maximum of 43.9% errors. There were more false positive judgments than false negative

judgments, suggesting that interviewers tended to err on the side of assuming sexual activity (West and Kreuter, 2015).

The initial cluster analysis provided evidence of two interviewers that could be considered outliers (Figure 2), with one interviewer citing neighborhood features in 55.41% of her justifications (the next highest percentage being 17.12%), and another interviewer citing health reasons for 43.14% of her justifications (the next highest percentage being 27.50%). After dropping these two interviewers, the second cluster analysis presented evidence of four unique clusters of interviewers based on scaled distances between the clusters (Figure 3); that is, there were in fact distinct groups of interviewers based on their justification tendencies. Descriptive statistics on the 14 variables for each cluster are shown in Table 3.

The results in Table 3 suggest that the first cluster of interviewers is largely defined by a tendency to notice living arrangements and housing unit characteristics. The second cluster is largely defined by references to appearance and personality, and a relatively large word count; interviewers in this cluster appeared to use the widest diversity of cues, including references to age and relationship status. The third cluster is primarily defined by references to relationship status and guesses / gut feelings, while the fourth cluster focuses primarily on age, occasionally referring to relationship status and household characteristics but hardly anything else.

We now consider differences among these clusters in terms of the various accuracy measures for the sexual activity judgments. Table 4 presents estimates of the fixed effects and variance components for the two multilevel logistic regression models fitted to each dependent accuracy measure.

The observational “strategy” clusters describing interviewers based on their justification tendencies are able to explain significant portions of the unexplained variance among interviewers in terms of observation accuracy. In the case of overall gross difference rates, Clusters 2 (use of a diversity of cues) and 3 (focus on relationship status, or “gut feelings”) have significantly reduced odds (odds ratio for Cluster 2 = $\exp(-0.40) = 0.67$, or 33% lower odds, and odds ratio for Cluster 3 = $\exp(-0.35) = 0.70$, or 30% lower odds, respectively) of making an error relative to Cluster 1. Furthermore, 36.2% of the unexplained variance among interviewers (when adjusting for the fixed effects of the area-level covariates) is accounted for by these “strategy” clusters.

In the case of false positive rates, Cluster 4 (focus primarily on age) had substantially increased odds of making a false positive error (odds ratio for Cluster 4 = $\exp(1.70) = 5.47$, or 447% higher odds), and the fixed strategy cluster effects explain about 15% of the unexplained variance in accuracy among interviewers. In the case of false negative rates, Clusters 2 and 3 once again have significantly reduced odds of making a false negative error relative to Cluster 1 (odds ratio for Cluster 2 = $\exp(-0.87) = 0.42$, or 58% lower odds, and odds ratio for Cluster 3 = $\exp(-0.99) = 0.37$, or 63% lower odds, respectively). We also note that Cluster 4 has reduced odds of making a false negative error, but the frequency with which this occurred was too small to prevent reliable estimation of this effect (see Table 3).

Collectively, the fixed cluster effects were able to explain nearly 44% of the unexplained variance in false negative rates among interviewers.

We therefore find consistent evidence of the observational strategies influencing the error properties of the sexual activity judgments in a significant manner, as described below:

- Focusing primarily on relationship status and gut feelings will result in the highest accuracy;
- Judging based primarily on age will result in systematic false positives;
- Considering a diversity of cues, including appearance, personality, and other external features, will also improve accuracy;
- Focusing on living arrangement and housing unit features is detrimental in terms of accuracy and results in systematic false negatives.

What might cause the remaining unexplained variance in accuracy among interviewers? All of the NSFG interviewers were female, and the vast majority were white, married, and had previous NSFG experience. For as many interviewers as possible (42 of 45), we extracted their overall years of interviewing experience, their age, and the number of children that they had from a voluntary interviewer survey (three of the 45 interviewers chose not to participate in the voluntary survey). NSFG managers examined these 42 values for each variable, and reported no evidence of measurement error given their knowledge of these particular interviewers. In exploratory analyses, we added fixed effects of these interviewer-level covariates to the “full” models in Table 4. For the gross difference rates, we found that interviewers with more kids had significantly reduced odds of making an error overall, with the effects of the strategy clusters remaining the same, and that the interviewer variance component was reduced to the point where it was no longer significantly greater than zero. For the false positive rates, we found that older interviewers had significantly reduced odds of making a false positive error, again further reducing the variance component. Finally, for the false negative rates, we found the older interviewers had significantly *increased* odds of making a false negative error; older interviewers appeared to err in the direction of no sexual activity. Relevant interviewer features therefore did seem to explain additional variance in accuracy, which is consistent with the existing literature (Sinibaldi et al., 2013; West and Kreuter, 2013; West and Kreuter, 2015) and could have training implications depending on the type of observation being collected.

Discussion

This study has demonstrated that the collection and analysis of open-ended justifications for the observations that field interviewers are often asked to record while conducting face-to-face surveys is feasible in practice. The analyses provide interesting insights into the observational strategies used by NSFG interviewers who make more accurate judgments regarding a specific respondent characteristic. With regard to our first research question, we found evidence of distinct clusters of interviewers based on the justifications that they tended to use for their judgments. This finding suggests that with only minimal guidance and training on the observation process provided by NSFG staff, different interviewers did in

fact use different observational strategies in the field when recording these types of judgments, consistent with our theoretical expectations. This finding certainly needs to be replicated in other survey contexts to further understand this phenomenon.

With regard to our second research question, the four clusters of interviewers identified based on the observational strategies evident in their justifications were found to vary significantly in terms of the error properties of this specific interviewer observation, when adjusting for the effects of other area- and interviewer-level covariates on accuracy. This finding suggests that variance in error rates on these types of observations may in fact be a function of varying observational strategies being employed in the field, which has important implications for standardized training on the process of recording interviewer observations (e.g., Dalhamer et al., 2012).

Our results were mainly consistent with theoretical expectations, where a focus on highly relevant cues (Funder, 1995) and more diversity in the cues used (Kazdin, 1977) was found to improve the error properties of the observations. Slightly contrary to theoretical expectations was the finding that a *combination* of focusing on a highly relevant cue (mention of relationship status during the screening interview) and guessing or “going on hunches” (expected to result in lower observation accuracy) was found to produce favorable error properties. A reasonable suggestion for practice would thus be to first try and identify highly relevant correlates of a feature being observed, and then go with general impressions or best guesses if those correlates are not readily available. Identification of these highly relevant features for a particular survey will require replications of this study, or at least discussions with interviewers who are found to produce highly accurate observations in a given survey. Either way, we hope that the methods described in this study will be used by other survey researchers to better understand effective observational strategies in other survey contexts.

Importantly, the different strategies used by interviewers to record the observations could also be reflective of the approaches that they take to recording the actual survey measurements in the interview. For example, an interviewer who, based on their background and expectations, uses age cues exclusively to justify their current sexual activity judgments may ultimately make comments or express opinions about risky sexual behavior in older or younger people during the actual interview. This could lead respondents to answer questions about current sexual activity in different (and possibly error-prone) ways, depending on statements made by the interviewer. A review of the literature on measurement of sexual behaviors in surveys (Fenton et al., 2001) suggests that interviewer gender can influence self-reports of sexual behavior (which would not be relevant in the NSFG, given that all interviewers are females), but also that establishment of rapport between the interviewer and the respondent can lead to reports of more frequent (and possibly exaggerated) sexual behavior. Interviewers may become more conversational and go off on tangents when rapport has been established, increasing the probability of interviewers communicating (either verbally or non-verbally) their expectations regarding the topic at hand (West and Blom, 2016), and this is when their “native” observational strategies might play a role in affecting measurement. If this were the case in practice, comparing the “strategy clusters” in

terms of observation accuracy may be misleading, but this requires future research (possibly taking advantage of CARI technologies for recording the survey interviews).

Alternatively, if older NSFG respondents tend to provide more socially desirable responses about sexual activity with respect to their age (Fenton et al., 2001), then in the group of interviewers that tended to use age exclusively in their justifications (Cluster 4 in this study), we may be estimating the error rates associated with this strategy incorrectly. For example, if 1) respondents between the ages of 20 and 49 tend to report being sexually active when they actually are not, 2) an interviewer tends to judge current sexual activity based on age (e.g., older respondents are more likely to be sexually active), and 3) the social desirability bias is larger for older individuals than younger individuals, what seem like correct judgments for older individuals may actually be incorrect at higher rates. This speculation would require additional research as well, and each of these possibilities speaks to the importance of having good validation data for establishing the accuracy of the interviewer judgments.

Setting aside these possibilities of a link between the observational strategy used and the accuracy of the respondent reports, the findings of this study have direct practical implications for future interviewer training. At a minimum, NSFG interviewers could be provided with verbal guidance about strategies to avoid (e.g., focusing on living arrangement) and strategies to employ (e.g., using a diversity of cues and attempting to detect some hint of relationship status in the screening interview) when recording these judgments. More generally, NSFG interviewers could watch brief videos of “staged” hypothetical screening interviews, where particular cues are mentioned (e.g., “My boyfriend isn’t home right now...”), and then be asked more generally about what judgments they might record and why. Thinking even more broadly about other surveys and other types of interviewer observations, some survey programs are starting to incorporate practice sessions for recording interviewer observations based on real photographs of housing units and neighborhoods into interviewer training (e.g., Dahlhamer, 2012; Stähli, 2010). Importantly, these sessions and the “correct” responses for what observations to record should be based on the type of empirical evidence generated in the present study.

There are several opportunities for future research in this area. First, the interviewer judgments in this study were only recorded for housing units where a screening interview was successfully completed (about 91% of housing units sampled for the NSFG). In general, the literature in this area would benefit from more case studies discussing the collection and analysis of “tailored” observations that are correlated with key survey variables and response propensity for *all* sampled units in a given survey. More evidence of such observations contributing to effective nonresponse adjustments would provide empirical support for the continued use of this practice. Second, the observations could only be validated using information provided by survey respondents. The possibility that observations on non-responding units may have suffered from reduced quality could not be considered in this study, and future studies would need to consider alternative sources of validation data to investigate this possibility further. Third, we spoke with a lawyer about the possibility that privacy regulations affecting data collection practices in Europe may eventually impact the ability of survey organizations to collect interviewer observations, and the lawyer indicated that interviewer observations are not an issue with respect to the current European privacy

laws. It will therefore be interesting to see if future debates about privacy protection make it more difficult to collect these types of observations, both in Europe and other countries.

Acknowledgments

Funding Funding for this research was provided by NIH Grant #1-R03-HD-075979-01-A1. The National Survey of Family Growth (NSFG) is carried out under a contract with the CDC's National Center for Health Statistics, Contract # 200-2010-33976. Funding comes from several agencies, including NCHS, NICHD, CDC, OPA and ACF. The views expressed here do not represent those of NCHS or the other funding agencies. We are indebted to Ziming Liao from the University of Michigan Undergraduate Research Opportunity Program (UROP) for his contributions to this work.

References

- Ambady N, Hallahan M, Conner B. Accuracy of Judgments of Sexual Orientation from Thin Slices of Behavior. *Journal of Personality and Social Psychology*. 1999; 77(3):538–547. [PubMed: 10510507]
- Babbie, ER. *The Practice of Social Research*. 9. Belmont, CA: Wadsworth /Thomson Learning; 2001.
- Baruch Y, Holtom BC. Survey response rate levels and trends in organizational research. *Human Relations*. 2008; 61(8):1139–1160.
- Beaumont JF. On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*. 2005; 31(2):227–231.
- Bethlehem, JG. Weighting nonresponse adjustments based on auxiliary information. In: Groves, R.Dillman, D.Eltinge, J., Little, R., editors. *Survey Nonresponse*. New York: Wiley; 2002. p. 275–287.
- Biemer PP. Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*. 2010; 74(5):817–848.
- Biener L, Garrett CA, Gilpin EA, Roman AM, Currivan DB. Consequences of Declining Survey Response Rates for Smoking Prevalence Estimates. *American Journal of Preventative Medicine*. 2004; 27(3):254–257.
- Brick JM, Williams D. Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *The ANNALS of the American Academy of Political and Social Science*. 2013; 645:36–59.
- Campanelli, P., Sturgis, P., Purdon, S. *Can you hear me knocking: An investigation into the impact of interviewers on survey response rates*. London: SCPR; 1997.
- Casas-Cordero C, Kreuter F, Wang Y, Babey S. Assessing the Measurement Error Properties of Interviewer Observations of Neighbourhood Characteristics. *Journal of the Royal Statistical Society (Series A)*. 2013; 176:227–250.
- Cull WL, O'Connor KG, Sharp S, Tang SS. Response rates and response bias for 50 surveys of pediatricians. *Health Services Research*. 2005; 40(1):213–226. [PubMed: 15663710]
- Curtin R, Presser S, Singer E. Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*. 2005; 69:87–98.
- Dahlhamer, J. New Observation Questions. Presentation at the 2013 NHIS Centralized Refresher Training and Conference; December 3–6, 2012; Hyattsville, MD. 2012. Available from the author (fzd2@cdc.gov)
- Das TH. Qualitative Research in Organizational Behaviour. *Journal of Management Studies*. 1983; 20:301–314.
- de Leeuw, E., de Heer, W. Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Chapter 3. In: Groves, RM., et al., editors. *Survey Nonresponse*. Wiley; 2002.
- Deming WE. On Errors in Surveys. *American Sociological Review*. 1944; 9:359–369.
- Everitt, BS., Landau, S., Leese, M., Stahl, D. *Cluster Analysis*. 5. 2011. Wiley Series in Probability and Statistics
- Feldman JJ, Hyman H, Hart CW. A Field Study of Interviewer Effects on the Quality of Survey Data. *Public Opinion Quarterly*. 1951; 15:734–761.

- Fenton KA, Johnson AM, McManus S, Erens B. Measuring Sexual Behavior: Methodological Challenges in Survey Research. *Sexually Transmitted Infections*. 2001; 77:84–92. [PubMed: 11287683]
- Forgas JP, Brown LB. Environmental and Behavioral Cues in the Perception of Social Encounters: An Exploratory Study. *The American Journal of Psychology*. 1977; 90:635–644.
- Funder DC. Errors and Mistakes: Evaluating the Accuracy of Social Judgment. *Psychological Bulletin*. 1987; 101:75–90. [PubMed: 3562704]
- Funder DC. On the Accuracy of Personality Judgment: A Realistic Approach. *Psychological Review*. 1995; 102:652–670. [PubMed: 7480467]
- Graham RJ. Anthropology and O.R.: The Place of Observation in Management Science Process. *The Journal of the Operational Research Society*. 1984; 35:527–536.
- Groves RM. Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*. 2006; 70(5):646–675.
- Groves RM, Lyberg L. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*. 2010; 74(5):849–879.
- Groves, RM., Wagner, J., Peytcheva, E. Use of Interviewer Judgments about Attributes of Selected Respondents in Post-Survey Adjustments for Unit Nonresponse: An Illustration with the National Survey of Family Growth. *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*; Salt Lake City, UT. 2007.
- Harris KJ, Jerome NW, Fawcett SB. Rapid Assessment Procedures: A Review and Critique. *Human Organization*. 1997; 56:375–378.
- Jones EE, Riggs JM, Quattrone G. Observer Bias in the Attitude Attribution Paradigm: Effect of Time and Information Order. *Journal of Personality and Social Psychology*. 1979; 37:1230–1238.
- Kazdin AE. Artifact, Bias, and Complexity of Assessment: The ABCs of Reliability. *Journal of Applied Behavior Analysis*. 1977; 10:141–150. [PubMed: 16795543]
- Kreuter F, Olson K, Wagner J, Yan T, Ezzati-Rice T, Casas-Cordero C, Lemay M, Peytchev A, Groves RM, Raghunathan TE. Using Proxy Measures of Survey Outcomes to Adjust for Survey Nonresponse. *Journal of the Royal Statistical Society, Series A*. 2010; 173(2):389–407.
- Kim Y, Choi YK, Emery S. Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician*. 2013; 67(3):171–182.
- Lepkowski JM, Mosher WD, Davis KE, Groves RM, Van Hoewyk J. The 2006–2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey. *National Center for Health Statistics. Vital and Health Statistics*. 2010 Jun.2(150)
- Lessler, J., Kalsbeek, W. Nonsampling Errors in Surveys. Wiley-Interscience; 1992. Nonresponse: Dealing with the Problem. Chapter 8 in.
- Little RJ, Vartivarian S. Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*. 2005; 31(2):161–168.
- Manderson L, Aaby P. An Epidemic in the Field? Rapid Assessment Procedures and Health Research. *Social Science Medicine*. 1992a; 35:839–850. [PubMed: 1411684]
- Manderson L, Aaby P. Can Rapid Anthropological Procedures Be Applied to Tropical Diseases? *Health Policy and Planning*. 1992b; 7:46–55.
- McCall GJ. Systematic Field Observation. *Annual Review of Sociology*. 1984; 10:263–282.
- McCulloch, SK., Kreuter, F., Calvano, S. Interviewer Observed vs. Reported Respondent Gender: Implications on Measurement Error. Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research; Chicago, IL. 5/14/2010; 2010.
- Millen, DR. Rapid Ethnography: Time Deepening Strategies for HCI Field Research. *Proceedings on DIS00: Designing Interactive Systems: Processes, Practices, Methods, and Techniques*; August 17–19, 2000; Brooklyn, NY. 2000. p. 280–286.
- Most SB, Scholl BJ, Clifford ER, Simons DJ. What You See is What You Get: Sustained Inattentive Blindness and the Capture of Awareness. *Psychological Review*. 2005; 112:217–242. [PubMed: 15631594]
- Patterson ML, Stockbridge E. Effects of cognitive demand and judgment strategy on person perception accuracy. *Journal of Nonverbal Behavior*. 1998; 22(4):253–263.

- Pickering, K., Thomas, R., Lynn, P. Testing the shadow sample approach for the English House Condition survey. Prepared for the Office of the Deputy Prime Minister by the National Centre for Social Research; London. July 2003; 2003.
- Punj G, Stewart DW. Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*. 1983; 20(2):134–148.
- Repp AC, Nieminen GS, Olinger E, Brusca R. Direct Observation: Factors Affecting the Accuracy of Observers. *Exceptional Children*. 1998; 55:29–36.
- Seidler J. On Using Informants: A Technique for Collecting Quantitative Data and Controlling Measurement Error in Organization Analysis. *American Sociological Review*. 1974; 39:816–831.
- Simons DJ, Jensen MS. The Effects of Individual Differences and Task Difficulty on Inattentive Blindness. *Psychonomic Bulletin and Review*. 2009; 16:398–403. [PubMed: 19293113]
- Sinibaldi J, Durrant GB, Kreuter F. Evaluating the Measurement Error of Interviewer Observed Paradata. *Public Opinion Quarterly*. 2013; 77:173–193.
- Sinibaldi J, Trappmann M, Kreuter F. Which is the Better Investment for Nonresponse Adjustment: Purchasing Commercial Auxiliary Data or Collecting Interviewer Observations? *Public Opinion Quarterly*. 2014; 78(2):440–473.
- Stähli, ME. Examples and Experiences from the Swiss Interviewer Training on Observable Data (neighborhood characteristics) for ESS 2010 (R5). Paper presented at the NC Meeting; Mannheim. 3/31/2011 – 4/1/2011; 2010.
- Stefanski LA, Carroll RJ. Covariate Measurement Error in Logistic Regression. *The Annals of Statistics*. 1985; 13(4):1335–1351.
- Tipping, S., Sinibaldi, J. Examining the trade off between sampling and targeted non-response error in a targeted non-response follow-up. Paper presented at the 2010 International Total Survey Error Workshop; Stowe, Vermont. June 15, 2010; 2010.
- Tolonen H, Helakorpi S, Talala K, Helasoja V, Martelin T, Prattala R. 25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey. *European Journal of Epidemiology*. 2006; 21:409–415. [PubMed: 16804763]
- Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. *Science*. 1974; 185:1124–1131. [PubMed: 17835457]
- Ward JH. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*. 1963; 58:236–244.
- West BT. An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth (NSFG). *Journal of the Royal Statistical Society (Series A)*. 2013a; 176:211–225.
- West, BT. The Effects of Error in Paradata on Weighting Class Adjustments: A Simulation Study. Chapter 15. In: Kreuter, F., editor. *Improving Surveys with Paradata: Making Use of Survey Process Information*. Hoboken, NJ: Wiley; 2013b. p. 361–388.
- West BT, Blom AG. Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*. 2017; 5:175–211.
- West BT, Kreuter F. Factors Impacting the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth (NSFG). *Public Opinion Quarterly*. 2013; 77:522–548.
- West BT, Kreuter F. A Practical Technique for Improving the Accuracy of Interviewer Observations of Respondent Characteristics. *Field Methods*. 2015; 27:144–162.
- West BT, Kreuter F, Trappmann M. 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*. 2014; 2:159–181.
- Williams, D., Brick, JM. Trends in U.S. Face-to-Face Household Survey Nonresponse and Level of Effort. *Journal of Survey Statistics and Methodology*. 2017. DOI: <https://doi.org/10.1093/jssam/smx019>
- Zhang, D., Lin, X. *Random Effect and Latent Variable Model Selection*. Springer; New York: 2008. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics; p. 19–36.

NSFG 7 Study - Contact Observations

Forms Answer Help

• After screening the household, you guessed that the selected respondent is not in an active sexual relationship with an opposite-sex partner. Why?

Respondent Questions

5	No
---	----

Rsex rel with opposite sex partner

CO_3 100100188011 12/14/2009 5:01:10 PM

Figure 1. NSFG interviewers entered open-ended justifications for their sexual activity judgments into the box labeled “Rsex rel with opposite sex partner.”

Figure 2.
Dendrogram showing results of initial cluster analysis, with evidence of two outliers (Interviewers 36 and 40).

Figure 3.
Dendrogram showing results of second cluster analysis, with evidence of four distinct groups of interviewers (based on rescaled cluster distances greater than 10).

Table 1

Descriptive statistics for the variables used in the cluster analysis.

	Mean	SD	Minimum	Maximum
Percentage of Justifications Mentioning:				
Relationship Status	43.25	13.70	17.86	75.00
Age	32.67	23.21	0.00	88.24
Living Arrangement	24.53	19.27	0.00	88.76
Housing Unit Characteristics	23.75	13.18	0.00	62.50
Guessing / Going on Hunches	12.10	17.62	0.00	82.14
Appearance	6.86	8.22	0.00	33.00
Occupation	4.43	5.82	0.00	25.00
Personality	3.88	4.70	0.00	17.82
Health	3.32	7.89	0.00	43.14
Neighborhood Characteristics	3.16	8.77	0.00	55.41
Conservative	1.69	2.90	0.00	12.50
Incorrect	1.01	1.81	0.00	8.70
Shyness	0.39	0.91	0.00	4.00
Mean Word Count	6.32	4.03	1.90	27.92

NOTE: $n = 45$ interviewers.

Table 2

Descriptive statistics for the interviewer-level error rates.

	Mean	SD	Minimum	Maximum
Gross Difference Rate	0.206	0.079	0.000	0.439
False Positive Rate	0.537	0.029	0.000	1.000
False Negative Rate	0.116	0.012	0.000	0.500

NOTE: $n = 45$ interviewers, with the exception of the false positive rates ($n = 43$).

Table 3

Descriptive statistics for interviewer-level justification tendencies and mean word counts within four distinct clusters of interviewers ($n = 43$ total).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	
Number of Interviewers	20	7	11	5	
Percentage of Justifications Mentioning:					Kruskal-Wallis χ^2 (df), p-value
Relationship Status	45.90 (11.10)	41.87 (7.04)	47.76 (18.43)	25.63 (5.85)	10.3 (3), $p = 0.016$
Age	23.02 (14.94)	49.96 (20.99)	21.03 (20.43)	57.52 (12.52)	17.2 (3), $p = 0.001$
Living Arrangement	37.62 (18.36)	20.96 (12.33)	11.11 (8.38)	2.06 (2.35)	26.1 (3), $p < 0.001$
Housing Unit Characteristics	28.60 (13.25)	25.11 (9.33)	13.04 (10.81)	26.23 (14.42)	9.0 (3), $p = 0.029$
Guess / Gut Feelings	5.95 (8.41)	3.96 (6.18)	33.93 (22.14)	1.55 (2.99)	17.0 (3), $p = 0.001$
Appearance	5.62 (4.01)	20.60 (9.67)	1.90 (3.89)	0.39 (0.54)	24.1 (3), $p < 0.001$
Occupation	5.59 (6.13)	5.73 (4.41)	1.33 (2.09)	0.00 (0.00)	14.7 (3), $p = 0.002$
Personality	3.85 (3.20)	8.94 (6.33)	1.09 (1.24)	0.27 (0.61)	17.0 (3), $p = 0.001$
Health	1.29 (1.61)	7.22 (10.32)	2.23 (4.62)	0.27 (0.61)	6.41 (3), $p = 0.093$
Neighborhood Characteristics	2.78 (4.71)	3.81 (3.44)	0.38 (1.08)	0.00 (0.00)	9.5 (3), $p = 0.023$
Conservative	2.68 (3.70)	1.75 (2.86)	0.59 (0.76)	0.00 (0.00)	5.94 (3), $p = 0.115$
Incorrect	1.29 (1.74)	0.81 (1.16)	0.24 (0.54)	2.29 (3.77)	4.06 (3), $p = 0.255$
Shyness	0.25 (0.55)	0.98 (1.41)	0.05 (0.16)	0.00 (0.00)	7.89 (3), $p = 0.048$
Mean Word Count	6.32 (2.48)	7.01 (1.58)	4.17 (1.31)	4.96 (2.29)	11.9 (3), $p = 0.008$
Gross Diff. Rate	0.247	0.191	0.168	0.171	
False Positive Rate	0.413	0.536	0.515	0.795	
False Negative Rate	0.196	0.087	0.070	0.006	

NOTES: Cells contain Mean (SD). The highest mean for each input variable is in **boldface**.

Table 4

Parameter estimates (logit scale) in multilevel logistic regression models indicating the relationships of selected covariates and interviewer “strategy” cluster membership with gross difference rates, false positive rates, and false negative rates ($n = 43$ interviewers for Models 1 and 3; $n = 41$ for Model 2).

	Model 1: Gross Difference Rates		Model 2: False Positive Rates		Model 3: False Negative Rates	
	Base Model	Including Strategy Clusters	Base Model	Including Strategy Clusters	Base Model	Including Strategy Clusters
Fixed Effects	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept	-1.57*** (0.56)	-1.82*** (0.54)	2.22 (1.68)	2.08 (1.68)	-2.35* (1.39)	-2.68** (1.23)
% Urban	0.10 (0.16)	0.17 (0.15)	0.05 (0.48)	<0.01 (0.48)	0.24 (0.42)	0.40 (0.36)
% Acc. Probs.	<0.01 (<0.01)	<0.01 (<0.01)	-0.01 (0.01)	-0.01 (0.01)	<0.01 (0.01)	<0.01 (0.01)
% Residential	-0.01 (<0.01)	-0.01 (<0.01)	<0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
% Non-English	-0.02 (0.02)	-0.01 (0.02)	0.04 (0.06)	-0.02 (0.06)	-0.06 (0.05)	0.12 (0.18)
% Spanish	0.02 (0.02)	0.01 (0.02)	-0.05 (0.05)	0.01 (0.06)	0.07 (0.04)	-0.12 (0.17)
% Safety Conc.	0.01 (<0.01)*	<0.01 (<0.01)	-0.02 (0.01)*	-0.01 (0.01)	0.02 (0.01)**	0.02 (0.01)*
% Many Units	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
% Phys. Imp.	-0.01 (0.01)	-0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	-0.02 (0.01)*	-0.02 (0.01)
% Female	<0.01 (0.01)	0.01 (0.01)	-0.04 (0.03)	-0.05 (0.03)	-0.01 (0.03)	0.01 (0.02)
Cluster 1: Living Arr. / Housing Unit Features		--		--		--
Cluster 2: Diversity of Cues		-0.40**		0.75		-0.87**

	Model 1: Gross Difference Rates		Model 2: False Positive Rates		Model 3: False Negative Rates	
	Base Model	Including Strategy Clusters	Base Model	Including Strategy Clusters	Base Model	Including Strategy Clusters
		(0.17)		(0.54)		(0.39)
<i>Cluster 3: Relationship Status / Hunches</i>		−0.35**		0.45		−0.99**
		(0.15)		(0.45)		(0.36)
<i>Cluster 4: Primarily Age</i>		−0.32		1.70**		−5.91
		(0.24)		(0.72)		(4.07)
Variance Components						
Interviewer Variance	0.058***	0.037***	0.904***	0.769***	0.732***	0.412***
	(0.029)	(0.022)	(0.333)	(0.289)	(0.243)	(0.149)
% Reduction in Interviewer Variance		36.2%		14.9%		43.7%

-- indicates reference category.

p <= 0.01,
**
p <= 0.05,
*
p <= 0.10.