Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2013)
176, *Part* 1, *pp.* 211–225

An examination of the quality and utility of interviewer observations in the National Survey of Family Growth

Brady T. West

University of Michigan, Ann Arbor, USA

[Received January 2011. Final revision January 2012]

Summary. Survey agencies have started to use interviewer observations collected on all sample units to adjust survey estimates for non-response. Ideally, these observations should be related to both response indicators and key survey variables. However, these observations are typically judgements that are made by the interviewers, making them potentially prone to measurement error. Presenting analyses of data from the National Survey of Family Growth in the USA, this study examines the quality and utility of these interviewer observations and considers the implications of measurement errors in these observations for the effectiveness of non-response adjustments.

Keywords: Error in auxiliary variables; Interviewer observations; Non-response adjustment; Paradata; Survey non-response

1. Introduction

Given declining response rates in surveys of nearly all formats 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) and rising costs of data collection, survey researchers increasingly rely on post-survey non-response adjustments to correct for potential non-response bias. Many of these adjustments require auxiliary variables that are available for both the respondents and the non-respondents in a given sample. Reductions of both the non-response bias and the variance in estimates are possible when these auxiliary variables are related to both the survey variables of interest and response indicators (Beaumont, 2005; Bethlehem, 2002; Groves, 2006; Kreuter *et al.*, 2010; Lessler and Kalsbeek, 1992; Little and Vartivarian, 2005). Unfortunately, auxiliary variables having these optimal properties are rare in practice (Kreuter *et al.*, 2010).

As a result, large survey research programmes have turned to the collection of paradata (Beaumont, 2005; Couper, 1998; Couper and Lyberg, 2005), or variables describing interviewer observations and other measurements about the survey data collection process, from both respondents and non-respondents (Kreuter *et al.*, 2010). A growing body of research has found associations of these paradata, and specifically interviewer observations, with contact indicators for sampled households (Blom *et al.*, 2011; Durrant *et al.*, 2011), response indicators (Durrant *et al.* (2012), Blom *et al.* (2011), Campanelli *et al.* (1997), Durrant *et al.* (2010), page 13, Durrant and Steele (2009), Groves and Heeringa (2006) and Lynn (2003)) and key survey variables (Kreuter *et al.*, 2010). However, interviewer observations are typically judgements

Address for correspondence: Brady T. West, Institute for Social Research, Survey Methodology Program, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106, USA.
E-mail: bwest@umich.edu

 0964–1998/13/176211

that are made by the interviewers, making them potentially prone to measurement error. Few published studies to date have examined the quality of these observations. In addition, the implications of errors in these observations for the effectiveness of post-survey non-response adjustments have yet to receive any research focus, and this study aims to address this gap in the literature.

This study analyses data from the recently completed seventh cycle of the National Survey of Family Growth (NSFG) and aims to address the following four research questions regarding selected interviewer observations in the NSFG.

(a) What are the measurement error properties of these interviewer observations?
(b) Are the interviewer observation variables associated with
    (i) response indicators and
    (ii) key survey variables?
(c) Do key survey estimates change when using interviewer observation variables in non-response adjustments in comparison with adjustments without these observations?
(d) How do measurement errors in the observations affect non-response adjustments?

The remainder of the paper is structured as follows. Section 2 reviews the existing literature on the quality of interviewer observations and the implications of reduced quality in the observations for non-response adjustments. Section 3 describes the NSFG data that are analysed in the paper. Section 4 details the statistical analyses that are used to answer each of the four research questions, and Section 5 presents the results of the analyses. Section 6 provides a summary of the findings and concludes with implications for practice and directions for future research in this area.

The programs that were used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Background

Motivated by previous attempts to use interviewer observations and other types of paradata for making non-response adjustments in the American National Election Studies (ANES), the European Social Survey and a national US transportation survey, Kreuter *et al.* (2010) demonstrated that interviewer observations that were collected in the NSFG were stronger correlates of key NSFG variables than similar paradata collected in the other surveys, and they showed that incorporating the observations into non-response adjustments led to moderate shifts in NSFG estimates. They called for additional research examining the measurement error properties of the observations and the effects of errors in the observations on adjusted survey estimates (Kreuter *et al.* (2010), page 17).

The associations of interviewer observations with response indicators and key survey variables may be attenuated by errors in the observations, which could in turn reduce the effectiveness of non-response adjustments based in part on the observations. The effects of measurement errors in auxiliary variables on the bias of estimated regression coefficients in linear regression models (e.g. Fuller (1987)) and logistic regression models (Stefanski and Carroll, 1985), which are often used for making non-response adjustments based on predicted response propensities, have been well established. The homogeneity of weighting classes constructed for non-response adjustments (in terms of response indicators and key survey variables) could therefore be adversely affected by the errors in these observations. For example, Lessler and Kalsbeek (1992), pages 189–190, showed that when respondents have higher means than non-respondents on a variable

of interest within each class, and the class level means and response rates are negatively correlated, adjusted estimates will have more bias than complete-case estimates. They concluded that

> 'these results warn us that it is possible to do more harm than good by using weighting class adjustments' (Lessler and Kalsbeek (1992), page 190).

Given the need for some form of validation data to evaluate interviewer observations, few published studies have been able to examine the errors directly (Campanelli *et al.*, 1997; Drury *et al.*, 1980; Hahn *et al.*, 1996; Pickering *et al.*, 2003). The rates of accuracy reported for the interviewer observations in these validation studies have ranged widely, suggesting that the quality of the observations can be quite low. Other studies have examined *indirect* indicators of error in the interviewer observations, including inter-interviewer reliability of the judgements (Alwin (2008), page 151, and Mosteller (1944)), interviewer variance in subjective judgements given interpenetrated assignments (Feldman (1951), page 743), interviewer problems with collecting the observations (Pickering *et al.*, 2003), variance in perceptions of respondent skin colour depending on interviewer race (Hill, 2002) and missing data rates for the observations (Lynn, 2003). Existing work suggests that the inter-interviewer reliability of *household* observations tends to be higher (Alwin (2008), page 151, and Mosteller (1944)) than that of judgements about *respondent* features.

No study to date has considered the implications of errors in interviewer observations for the effectiveness of non-response adjustments based in part on the observations. Steiner *et al.* (2011) considered observational studies, where propensity score modelling is often used to reduce possible selection bias in treatment effects. They demonstrated that when the covariates that are used in propensity score models are related to outcomes of interest but have reduced reliability the ability of the covariates to reduce selection bias will be significantly reduced. Biemer *et al.* (2012) studied sources of error in interviewer-recorded numbers of call attempts in the National Survey of Drug Use and Health, and the bias that these errors can introduce in non-response adjustments based on a *call-back model* (see Biemer *et al.* (2010), for a description of this model). Biemer *et al.* (2012) showed through simulation studies that when error rates for the call records vary depending on the disposition of a sample case (for example refusals have higher error rates compared with completed cases) the bias in estimated proportions based on the call-back models can be substantial. The present study aims to build on this recent work by examining how measurement errors in interviewer observations affect weighting class adjustments for non-response.

3. National Survey of Family Growth data

Data collected during cycle 7 of the NSFG (July 2006–June 2010) were analysed in this study. Initial screening interviews are necessary in the NSFG to determine the eligibility of individuals in randomly selected households, given that the target population is non-institutionalized males and females between the ages of 15 and 44 years living in the USA. Additional details on the design of the NSFG, which has a primary goal of collecting 'nationally representative data on factors affecting birth and pregnancy rates, family formation, and the risks of HIV and other STDs', can be found elsewhere (Groves *et al.*, 2009). The present study analyses six variables containing interviewer observations, two of which also have validation data available.

NSFG interviewers (all of whom are female) are typically assigned to work in a single primary sampling unit. Before the first attempted screening interview with a randomly selected household in their primary sampling unit, interviewers were instructed to locate the household and

to estimate whether it contained any children under the age of 15 years (the first interviewer observation variable). In the data set that was constructed for analysing the amount of error in these observations (research question (a)), there were a total of 58225 observations on the presence of children reported by 116 interviewers. For this first interviewer observation variable, completed household roster information was available to determine whether children under the age of 15 years were actually present in the household. For this study, completed household rosters were assumed to be correct, and care was taken to ensure that there were no errors in linking data sets containing household roster information and the interviewer observations.

Immediately after the completion of the screening interview and the selection of an eligible person from a household for the main NSFG interview, interviewers were asked to estimate whether the person selected was in a sexually active relationship with an opposite sex partner (the second interviewer observation variable). There were a total of 22669 judgements of sexual activity reported by 113 interviewers, for which actual survey information on sexual activity was also available from the computer-assisted personal interviewing portion of a *completed* main interview (for research question (a)).

The fact that 'true' values for current sexual activity were available for respondents to the main interview only prevents analyses of the relationships of the true current sexual activity with a response indicator, outside simulation studies. Possible errors in the respondent reports of sexual activity, assumed to be 'truth', were not considered in this study. Care was once again taken to prevent any errors in record linkage.

This study therefore considers two interviewer observation variables from the NSFG with available validation data:

(a) the estimate of whether children under the age of 15 years were present in a sampled household, collected before the screening interview, with 58225 estimates that could be validated by using household roster information;
(b) the judgement of whether a respondent who was randomly selected from the screening interview was currently in a sexually active relationship, with 22669 judgements that could be validated by using respondent reports from the main interview.

There were a total of 25451 completed screening interviews resulting in the selection of an eligible person. These two interviewer observations were available for each of these completed interviews, and the person selected either completed or did not complete the main interview (enabling response propensity modelling for research question (b)). True values on current sexual activity were not available for those people who did not complete a main interview, preventing validation of the current sexual activity judgements in these cases.

This study analyses five variables measured for main interview respondents:

(a) a binary indicator of whether the respondent had ever been married;
(b) a binary indicator of whether the respondent had ever cohabitated with a partner;
(c) the number of sexual partners in the past year;
(d) for males, a count of biological children;
(e) for females, parity, or the number of live births.

Male and female respondents to the main interview were coded as being sexually active if reporting one or more opposite sex partners in the past 12 months. Female respondents were also asked about having a current opposite sex partner, and this measure was used to indicate being sexually active if no information was available on the number of partners in the past 12 months. Measurement error was certainly possible for these variables, and error rates may have differed

for males and females, but implications of these errors are left to future research. Sampling weights and sampling error codes that were provided by NSFG staff for each of the respondents (Lepkowski *et al.*, 2010) were used for unbiased estimation of selected parameters and design-based estimation of the standard errors of the parameter estimates (research questions (b) and (c)).

4. Statistical analyses

Four separate statistical analyses were performed to address the first three research questions. First (research question (a)), unweighted κ-statistics were used to examine overall agreement of the interviewer judgements with the available validation data. Second (research question (b)), three multilevel logistic regression models were fitted to an indicator of response to the main interview (conditional on a completed screening interview), with random interviewer effects accounting for the clustering of households within interviewers. These models were fitted to the data set of 25451 successful screening interviews with interviewer observations available, and included

(a) the base NSFG sampling weight (to make the sampling uninformative with respect to the models),
(b) a series of auxiliary variables measured at the household level and
(c) two interviewer level characteristics (indicators for a non-white interviewer and a bilingual interviewer).

The first model did not include any interviewer observations. The second model added interviewer observations that could *not* be validated as predictors (noting physical impediments to the household, estimated probabilities of a main interview being completed, noting whether all housing units in a sampled area segment are residential and noting safety concerns in the segment), to assess their additional contribution to the model (via comparison of generalized χ^2-statistics; SAS Institute (2011)). The third model added the interviewer judgements of current sexual activity and presence of children as predictors, to analyse their independent ability to predict the response to the main interview. These three models were also refitted treating the interviewer effects as fixed (and excluding fixed effects of interviewer level covariates), enabling comparisons of rescaled pseudo-R^2-values for the three models (Nagelkerke, 1991). The three models were fitted by using residual pseudolikelihood estimation in SAS procedure `GLIMMIX`, and the variance of the random interviewer effects was tested against zero by using likelihood ratio tests, as described in Zhang and Lin (2008). Predicted response propensities were computed for each responding case on the basis of the estimated parameters in a given model, including the empirical best linear unbiased predictions of the random interviewer effects. To minimize additional variance in the estimates that was introduced by variance in the non-response adjustments, these response propensities were then grouped into deciles (or weighting classes), and the inverse of the unweighted proportion of responding cases within each of the 10 weighting classes was used to adjust the base sampling weights for non-response (Little, 1986).

Third (research question (b)), considering main interview respondents only, multilevel logistic and linear regression models were fitted to the five variables of interest by using SAS `PROC GLIMMIX`. Square-root transformations were applied to the three count variables (the number of partners in the past year, parity for females and the number of biological children for males) to stabilize variance in these responses (Faraway (2005), page 58). These models included the two interviewer observations that could be validated, along with the same covariates (including the base sampling weight and the other interviewer observations) and random

interviewer effects that were used in the third response propensity model. These analyses assume that the associations of the survey variables with the interviewer observations are the same for both respondents and non-respondents. Differential error rates for the observations for respondents and non-respondents (e.g. Matsuo *et al.* (2010), pages 35–43), which have been shown to affect non-response adjustments (Biemer *et al.*, 2012), may lead to different associations. The available NSFG data did not permit testing this assumption, because only the respondents had information available on both the interviewer observations and the survey variables.

Fourth (research question (c)), design-based estimates of means and percentages on the five survey variables were computed by using

(a) the base sampling weights,
(b) non-response-adjusted base weights with weighting class adjustments *excluding* the two interviewer judgements under study and
(c) non-response-adjusted base weights with weighting class adjustments *including* the two interviewer judgements.

These non-response adjustments assume that the NSFG non-respondents are missing at random (see Little and Rubin (2002), section 1.3), conditional on the weighting classes. Taylor series linearization (Wolter (2007), chapter 6) was used to estimate variances of the weighted means, accounting for the complex sample design of the NSFG and ignoring negligible finite population corrections (based on the large NSFG target population). This approach assumes that the non-response adjustments applied to the base weights are *fixed* quantities, which could lead to underestimation of the variance (Valliant, 2004). When base weights are multiplied by estimated response propensities, weighted estimates of means become ratios of *non-linear* functions of the random-sample inclusion indicators and response indicators, and this complicates the Taylor series linearization approach. In this setting, incorporating variance in the stochastic non-response adjustments is more straightforward when using replicated variance estimators (Valliant, 2004). Taylor series linearization was used in the present study as software is not readily available that implements the more appropriate replicated estimators, and variance estimation was not the main focus of this study.

Weighted estimates computed by using the same sample but different weights were compared statistically by 'stacking' three different versions of the same data file (with only the weights differing), and constructing confidence intervals for the differences in the means that incorporated the covariances of the estimates. These analyses were performed by using the `svy: mean,over()` command in conjunction with the `lincom` post-estimation command in Stata 11.2 (StataCorp, 2009).

Finally, Section 5.4 describes the simulation study that was used to address research question (d), examining the implications of the errors in the two NSFG interviewer observations for the effectiveness of subsequent non-response adjustments.

5. Results

This section reports results of the analyses for each of the four research questions in turn.

5.1. What are the measurement error properties of these interviewer observations?

Table 1 shows that 72.3% (i.e. 59.94% + 12.36%) of the interviewer judgements on presence of children were accurate, on the basis of completed household rosters. The unweighted κ-statistic computed for these data ($\kappa = 0.285$; 95% confidence interval $\mathrm{CI} = 0.276, 0.293$) suggests 'fair'

Table 1. Case counts and overall percentages indicating the error properties of interviewer judgements regarding the presence of children under the age of 15 years in selected households (NSFG, cycle 7)†

Interviewer judgement: children age < 15 years	*Household roster indicator: children age < 15 years*		*Total*
	No	*Yes*	
No	34898 (59.94%)	9028 (15.51%)	43926 (75.44%)
Yes	7103 (12.20%)	7196 (12.36%)	14299 (24.56%)
Total	42001 (72.14%)	16224 (27.86%)	58225 (100.00%)

†$\kappa = 0.285$; 95% CI (0.276, 0.293).

Table 2. Case counts and overall percentages indicating the error properties of interviewer judgements of current sexual activity among respondents (NSFG, cycle 7)†

Interviewer judgement: selected R sexually active	*Main NSFG interview: selected R sexually active*		*Total*
	No	*Yes*	
No	2230 (9.84%)	2081 (9.18%)	4311 (19.02%)
Yes	2912 (12.85%)	15446 (68.14%)	18358 (80.98%)
Total	5142 (22.68%)	17527 (77.32%)	22669 (100.00%)

†$\kappa = 0.334$; 95% CI (0.319, 0.349).

agreement of the judgements with the household rosters, per guidelines of Landis and Koch (1977), page 165. The false positive rate was 0.169 (7103 /42 001) whereas the false negative rate was 0.557 (9028/16 224), suggesting that detecting a child was a more difficult task than noting that no children were present.

The overall accuracy for judgements of current sexual activity approached 78% (Table 2), considering main interview respondents only ($\kappa = 0.334$; 95% CI = 0.319, 0.349; also considered 'fair' agreement per Landis and Koch (1977)). In contrast with the housing unit observations on the presence of children, the false positive rate for the sexual activity judgements was much higher (0.566) than the false negative rate (0.119), suggesting that judgement of sexual activity was much more difficult for people who were not sexually active.

5.2. Are the interviewer observation variables associated with response indicators and key survey variables?

Table 3 presents estimates of adjusted odds ratios in the three multilevel logistic regression models predicting propensity to respond to the main NSFG interview, conditional on a completed screening interview (see Section 4 for details). Model 1 includes only household and interviewer level covariates, whereas models 2 and 3 add the interviewer observation variables as covariates.

The results in Table 3 indicate that the interviewer observations are making substantial contributions to the NSFG response propensity model, as evidenced by the reductions in the

Table 3. Main interview response propensity modelling results, showing adjusted relationships of NSFG interviewer observation variables with response indicators†

Interviewer observation variable	*Estimated odds ratios (95% CIs) for the following models:*		
	Model 1	*Model 2*	*Model 3*
Interviewer notes physical impediments to household		0.985 (0.851, 1.139)	0.986 (0.852, 1.142)
Interviewer estimates high main interview probability		0.565 (0.479, 0.668)	0.559 (0.473, 0.661)
Interviewer estimates medium main interview probability		0.326 (0.260, 0.341)	0.295 (0.257, 0.339)
Interviewer estimates low main interview probability		0.093 (0.081, 0.106)	0.093 (0.081, 0.106)
Interviewer does not report main interview probability		Reference	Reference
Interviewer notes all housing units in segment residential		0.997 (0.902, 1.101)	0.999 (0.904, 1.104)
Interviewer notes safety concerns in segment		1.026 (0.915, 1.151)	1.018 (0.907, 1.142)
Interviewer estimates respondent sexually active			1.923 (1.707, 2.166)
Interviewer estimates children under 15 years in household			1.184 (1.064, 1.317)
Sample size	25451	25451	25451
Estimated variance of random interviewer effects (likelihood ratio test p-value)	0.261 ($p < 0.001$)	0.257 ($p < 0.001$)	0.274 ($p < 0.001$)
Generalized χ^2-statistic (SAS Institute, 2011)	24763.98	23005.99	22327.04
Pseudo-R^2 (treating interviewer effects as fixed)	0.271	0.357	0.365

†All three models controlled for additional household and interviewer level covariates, including the base sampling weight, number of calls, number of contacts, a black respondent indicator, indicators for different data collection quarters, age of selected respondent, an urban primary sampling unit indicator, a single-person household indicator, non-white interviewer, bilingual interviewer, census regions, sampling segment domains and a second-phase sample indicator (see Lepkowski *et al.* (2010) for additional details on these covariates). Estimated odds ratios for these covariates are not shown but are available on request. Models including fixed interviewer effects (enabling computation of pseudo-R^2-values) did not include the interviewer level covariates (non-white interviewer and bilingual interviewer).

generalized χ^2 fit statistics (SAS Institute, 2011) and increases in pseudo-R^2-values (treating interviewer effects as fixed) when the observations are added to model 1. Specifically, interviewer estimates of the probability that a main interview will be completed are strongly predictive of the response indicator, with cases having missing estimates (many of which are missing because of main interviews that are completed immediately following the screening interview) and cases with high estimated probabilities having higher relative probabilities of responding.

The two observations of interest are also significant predictors of the response indicator when controlling for all other auxiliary variables and interviewer observations (model 3), but they do not result in the same large improvement in model fit. Households that were estimated to have children under 15 years of age and selected respondents estimated to be sexually active had significantly higher probabilities of completing the main NSFG interview when adjusting for the other covariates. Unfortunately, we cannot examine the contributions of the true auxiliary variables measuring the presence of children and current sexual activity to the response

propensity model as both variables were only available for main interview respondents. Finally, we note evidence of substantial interviewer variance in response propensities (even after accounting for the interviewer level predictors of being non-white or bilingual), indicating that predicted response propensities may have to account for empirical best linear unbiased predictors of the random interviewer effects in these models.

Table 4 displays, for respondents only, adjusted estimates of the relationships of the two interviewer observation variables with the five key NSFG variables, alongside adjusted estimates of the relationships of the true auxiliary variables with each variable. These estimates were obtained by fitting the multilevel linear and logistic regression models that were described in Section 4 to the five survey variables.

The interviewer observation variables tended to have strong associations with the five key survey variables (fourth and sixth columns) when adjusting for the other auxiliary variables, including the base sampling weights (which were significantly related to three of the five survey variables, indicating informative sampling). However, it appears that the errors in the observations are severely attenuating the relationships of the two true auxiliary variables with the survey variables (fifth and seventh columns). Notably, there are large differences in the R^2- or pseudo-R^2-values for the fitted models when using the interviewer observations rather than the true values (third column), especially for the number of sexual partners in the past year. This finding suggests that use of the interviewer observations may be limiting the collective predictive power of these auxiliary variables (and thus the effectiveness of non-response adjustments including these two auxiliary variables). Finally, likelihood ratio tests for the variances of the random interviewer effects in these 10 models (which are not shown in Table 4) were significant for nine of the 10 models (with the exception of the model including true values of current sexual activity), suggesting substantial interviewer variance in responses even after accounting for a large number of auxiliary variables (including interviewer level covariates).

Table 4. Adjusted estimates of regression parameters for the two interviewer judgements as predictors of five key NSFG variables, contrasted with estimates by using the true auxiliary variables as predictors instead (NSFG respondents only)†

NSFG variable	*n*	*Pseudo-R^2-values*	*Interviewer judgement: children under 15 years*	*Household roster: children under 15 years*	*Interviewer judgement: sexual activity*	*Respondent report: sexual activity*
Ever been married	22682	0.523/0.571	0.20‡	0.80‡	1.17‡	1.95‡
Ever cohabitated	22682	0.283/0.364	0.13‡	0.32‡	0.98‡	1.99‡
Number of biological children (males only)	10403	0.369/0.466	0.20‡	0.48‡	0.22‡	0.43‡
Number of sexual partners in past year	21008	0.054/0.605	−0.01	−0.05‡	0.26‡	1.17‡
Parity (number of live births) (females only)	12279	0.418/0.489	0.22‡	0.73‡	0.23‡	0.32‡

†Parameter estimates for other covariates listed in the Table 3 notes are not shown for each dependent variable in the first column. Pseudo-R^2-values are computed from models with interviewer effects treated as fixed and interviewer level covariates omitted (as in Table 3). The first pseudo-R^2-value is for the model using the interviewer judgements, and the second value is for the model using the true values. Parameter estimates and tests of significance are based on models including random interviewer effects.

‡$p < 0.001$.

Table 5. Effects of alternative non-response adjustments on NSFG estimates†

Estimate (sample size)	*Base weights only*	*Non-response-adjusted base weights without the two observations*	*Non-response-adjusted base weights with the two observations*
% ever married ($n = 22682$)	49.527 (1.002)	50.005 (1.179)	49.922 (1.203)
% ever cohabitated ($n = 22682$)	49.533 (1.314)	49.881 (1.449)	49.782 (1.473)
Mean number of partners in past year ($n = 21008$)	1.130 (0.015)	1.130 (0.016)	1.128 (0.016)
Males: mean number of biological children ($n = 10403$)	1.266 (0.054)	1.304 (0.083)	1.300 (0.083)
Females: mean parity ($n = 12279$)	1.293 (0.041)	1.280 (0.043)	1.276 (0.044)

†Linearized standard errors are reported in parentheses. These estimates do not incorporate post-stratification factors or imputations of missing values and do not represent final estimates based on NSFG cycle 7.

5.3. Do key survey estimates change when using interviewer observation variables in non-response adjustments in comparison with adjustments without these observations?

Table 5 presents estimates of percentages or means (including linearized estimates of standard errors) on the five key NSFG variables, using the three alternative weights that were described in Section 4.

When the two interviewer observation variables that could be validated were included in the weighting class adjustments (fourth column), the mean for parity was found to be significantly lower than both the unadjusted mean using base weights only (95% CI for difference 0.0002, 0.0335) and the mean based on non-response adjustments excluding the two observations (95% CI for difference 0.0001, 0.0078). No other changes in means were significant. In general, including the two interviewer observation variables in the non-response adjustments did not result in substantial shifts in the total population or subgroup (male or female) estimates.

5.4. How do measurement errors in the observations affect non-response adjustments?

A simulation study was performed using real NSFG data to address this fourth research question. A hypothetical population was defined by $N = 10561$ female respondents to the main NSFG interview. The data set for this population included both interviewer judgements of current sexual activity and actual reports of current sexual activity from the main NSFG interview. Two NSFG variables measuring parity and number of partners in the past year were also included in the population data file for the simulations, and only females with complete data were included.

In each of six simulations (three weighting schemes for each survey variable), 1000 simple random samples of size $n = 500$ were selected from this hypothetical population. Unit non-response was simulated for each of the 1000 samples on the basis of the following logistic regression model, motivated by actual NSFG outcomes (see Table 3):

$$\Pr(\text{response}_i) = \frac{\exp(\text{report.sexually.active}_i)}{1 + \exp(\text{report.sexually.active}_i)}.$$

A sampled case denoted by i had values on the two survey variables deleted if a random draw from a uniform(0, 1) distribution was greater than or equal to the probability computed above. The simulated probability of response was thus a function of the *reported* sexual activity for case i (1, yes; 0, no), and not the interviewer judgement.

Table 6. Results of the simulation study, showing the empirical performance of estimators with non-response adjustments based on either respondent reports of current sexual activity or interviewer judgements of current sexual activity

NSFG variable	*Non-response adjustment method*	*Auxiliary variable for non-response adjustment*	*Empirical bias (relative %)*	*Empirical RMSE*	*95% CI coverage*
Parity	None		0.0064 (0.47%)	0.0755	0.953
(true mean	Weighting classes	Self-reported sexual activity	−0.0027 (−0.20%)	0.0762	0.948
= 1.3743)		Interviewer judgement of sexual activity	0.0085 (0.62%)	0.0714	0.968
Partners in	None		0.0319 (2.67%)	0.0566	0.904
past year (true	Weighting classes	Self-reported sexual activity	0.0012 (0.10%)	0.0453	0.960
mean = 1.1626)		Interviewer judgement of sexual activity	0.0326 (2.80%)	0.0573	0.906

In four of the six simulations, a simple weighting class adjustment for non-response was performed. Two weighting classes defined by categories for either the true respondent report of current sexual activity or the interviewer judgement were formed, and the inverse of the proportion of respondents within each class was used as a non-response adjustment weight. Two simulations computed non-response-adjusted estimates of mean parity for each sample by using the two alternative auxiliary variables, whereas two simulations computed non-response-adjusted estimates of the mean number of partners in the past year by using the alternative adjustments. Finally, two simulations computed complete-case estimates of means for parity and number of partners in the past year for each sample.

Since the means for the two survey variables are known, the following outcomes were examined for each of the six simulations:

(a) the empirical bias of the estimate (in terms of a percentage bias relative to the known mean);
(b) the empirical root-mean-squared error (RMSE) of the estimate;
(c) 95% confidence interval coverage of the estimate.

Standard errors of weighted estimates were computed by using Taylor series linearization, and confidence intervals were computed by assuming normally distributed estimates. Table 6 summarizes the results of this simulation study.

Table 6 shows that the use of interviewer judgements of current sexual activity as an auxiliary variable when constructing the non-response adjustments (the third and sixth rows) attenuates potential reductions in bias when using the weighting class adjustment method, relative to adjustments using the true self-reported values of current sexual activity (the second and fifth rows). This is especially true for survey variables having a stronger relationship with the auxiliary variable in question. In the hypothetical population, the correlation of the true current sexual activity with parity was only 0.072, whereas the correlation of true current sexual activity with the number of partners in the past year was 0.396. The adjusted estimate of the mean number of partners in the past year is most severely affected by using the error prone interviewer observations for non-response adjustment.

When using the interviewer judgements to define the two weighting classes, the bias of the resulting estimates is similar to that found when analysing the complete cases only under the

defined non-response mechanism (the first and fourth rows). The positive empirical bias in the complete-case estimator for the mean number of partners in the past year (2.67% relative bias) actually becomes *larger* when using non-response adjustments based on the error prone interviewer judgements (2.80% relative bias), which is consistent with the theoretical possibility that was suggested by Lessler and Kalsbeek (1992), page 190. We find this result because

(a) respondents tended to have a higher mean number of partners in the past year than non-respondents in the two classes formed by the interviewer judgements,
(b) the mean number of partners in the past year was actually *higher* in the weighting class defined by those judged to be *not* sexually active by the interviewers and
(c) the response propensity was lower in the weighting class mentioned in (b).

There is also evidence of higher empirical RMSEs in the estimates (compared with the complete-case estimators) when using the interviewer judgements, in contrast with the lower empirical RMSE in the estimates that was found when using the true values. Given that the MSE (RMSE squared) is defined as variance plus bias squared, the RMSEs in Table 6 indicate that the variances of the three estimators for mean parity are 0.0057, 0.0058 and 0.0050 whereas the variances of the three estimators for the number of sexual partners are 0.0022, 0.0021 and 0.0022. These results suggest that differences in the MSEs of the estimators are largely being driven by the bias that is introduced by the interviewer judgements, rather than the variance. There is also evidence in Table 6 that confidence interval coverage may be affected in a negative manner by the use of the error prone interviewer judgements for non-response adjustments, especially for survey variables having a stronger relationship with the variables that are judged by the interviewers.

6. Discussion

This study addressed four research questions regarding the quality and utility of interviewer observations in the NSFG. The first research question concerned the error properties of the observations. Two interviewer observations that could be validated by using household roster information and survey data had rates of accuracy around 70–80%. Errors on a person level judgement (current sexual activity) were largely false positive errors, whereas errors on a household level judgement (presence of children) were largely false negative errors. The present study adds to the existing literature by suggesting that errors in specific observations may be systematic rather than variable.

The second question concerned the associations of the interviewer observation variables with a main interview response indicator and five key NSFG variables. The two observation variables were found to have significant associations with both the main interview response indicator and the five key variables when adjusting for a variety of auxiliary variables and other interviewer observations. However, relationships of the two true auxiliary variables with the five survey variables were severely attenuated by the errors in the observations, as expected from existing theory. These findings suggest that the effectiveness of non-response adjustments based in part on error prone interviewer observations may be limited by their decreased ability to predict key survey variables.

The third question concerned whether including the interviewer observation variables in non-response adjustments shifted estimates of population means and percentages for the five NSFG variables. Weighting class adjustments incorporating the two interviewer observations shifted the estimated means and percentages slightly relative to adjustments without the observations, with a significant shift observed for estimates of mean parity (for females). These findings

were consistent with existing literature showing that non-response adjustments incorporating interviewer observations do not tend to shift estimates substantially (Kreuter *et al.*, 2010), and the answers to the first two research questions provide possible explanations for the reduced effectiveness of the adjustments.

Given that associations of the true auxiliary variables with the five NSFG variables were shown to be severely attenuated when using the error-prone interviewer observations in their place, the fourth research question examined whether the lack of a shift in the estimated means and percentages may be due to the errors in the observations. A simulation study supported existing theory, showing that the use of interviewer observations instead of true values on auxiliary variables to form weighting classes for non-response adjustments can attenuate potential reductions in bias, and, in the case of auxiliary variables having stronger relationships with key survey variables, lead to more bias in estimates than complete-case analyses.

The findings that were summarized above suggest that excessive error levels in interviewer observation variables may limit the effectiveness of non-response adjustments based in part on those variables. Survey researchers collecting interviewer observations with post-survey non-response adjustments in mind should therefore consider

(a) design strategies for improving the quality of the observations (e.g. West (2010)), especially given the systematic nature of the errors that was found in this study, and
(b) estimation techniques that (given validation data for respondents only) mitigate the effects of errors in the interviewer observation variables on non-response adjustments (e.g. West and Little (2011)).

Kott (2006), section 6, described a calibration method of using variables observed for respondents only to perform non-response adjustments, which may prove important in this context. Survey statisticians could also replace particularly error prone interviewer judgements of variables eventually measured in a survey with 'smoothed' predictions, based on auxiliary predictors of eventual respondent reports that are available at the time of a judgement (West, 2010).

The present study was largely limited by its focus on only two interviewer observations in one large national survey in the USA. Similar detailed investigations of both the quality of interviewer observations and the implications of poor quality for the effectiveness of alternative non-response adjustments are certainly needed in other survey contexts. The availability of true values for auxiliary variables being approximated with interviewer observations for *non-respondents* (in addition to respondents) would also enable study of the possible attenuating effects of errors in the observations on response propensity models. High quality administrative records may prove useful in this regard.

There are many avenues for future research in this area. First, analyses of trends in accuracy of observation over the life of a data collection are needed to see whether interviewers improve with more experience, which has important training implications (e.g. Stähli (2011)). Second, multilevel modelling techniques could be used to identify respondent and interviewer level covariates influencing the accuracy of interviewer observations, and knowledge of these covariates could in turn inform design strategies that are aimed at improving the quality of the observations. Third, the utility of interviewer estimates of the probability that a main interview will be completed for non-response adjustments should be studied carefully. Fourth, future simulation studies could extend the present study by considering subpopulation estimates (as opposed to the total sample estimates that were considered here) as well as different types of non-response adjustments, including calibration estimation (e.g. Särndal and Lundström (2010)). Finally, the implications of using error prone auxiliary variables for other survey

methodologies, including responsive survey designs (Groves and Heeringa, 2006), stratified sampling, and model-based imputation approaches, need future research attention.

Acknowledgements

The author acknowledges the helpful feedback and commentary from the Joint Editor, the Associate Editor, two reviewers and Mick Couper. The 2006–2010 NSFG was carried out under a contract with the Centers for Disease Control and Prevention's National Center for Health Statistics (contract 200-2000-07001).

References

Alwin, D. F. (2008) *Margins of Error: a Study of Reliability in Survey Measurement*. Hoboken: Wiley.
Baruch, Y. and Holtom, B. C. (2008) Survey response rate levels and trends in organizational research. *Hum. Relns*, **61**, 1139–1160.
Beaumont, J.-F. (2005) On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Surv. Methodol.*, **31**, 227–231.
Bethlehem, J. (2002) Weighting nonresponse adjustments based on auxiliary information. In *Survey Nonresponse* (eds R. Groves, D. Dillman, J. Eltinge and R. Little), pp. 275–287. New York: Wiley.
Biemer, P. P., Chen, P. and Wang, K. (2011) Errors in the recorded number of call attempts and their effect on nonresponse adjustments using callback models. To be published.
Biemer, P. P., Wang, K. and Chen, P. (2010) Using call-back data to adjust for nonignorable nonresponse: results of an empirical study. *Jt Statist. Meet.*, *Vancouver, May*.
Biener, L., Garrett, C. A., Gilpin, E. A., Roman, A. M. and Currivan, D. B. (2004) Consequences of declining survey response rates for smoking prevalence estimates. *Am. J. Preventv. Med.*, **27**, 254–257.
Blom, A. G., de Leeuw, E. D. and Hox, J. J. (2011) Interviewer effects on nonresponse in the European Social Survey. *J. Off. Statist.*, **27**, 359–377.
Campanelli, P., Sturgis, P. and Purdon, S. (1997) *Can You Hear Me Knocking: an Investigation into the Impact of Interviewers on Survey Response Rates*. London: Social and Community Planning Research.
Couper, M. P. (1998) Measuring survey quality in a CASIC environment. In *Proc. Jt Statist. Meet.*, *Dallas*. Alexandria: American Statistical Association.
Couper, M. P. and Lyberg, L. (2005) The use of paradata in survey research. In *Proc. 55th Sess. International Statistical Institute, Sydney*. Voorburg: International Statistical Institute.
Cull, W. L., O'Connor, K. G., Sharp, S. and Tang, S. S. (2005) Response rates and response bias for 50 surveys of pediatricians. *Hlth Serv. Res.*, **40**, 213–226.
Curtin, R., Presser, S. and Singer, E. (2005) Changes in telephone survey nonresponse over the past quarter century. *Publ. Opin. Q.*, **69**, 87–98.
Drury, T. F., Moy, C. S. and Poe, G. S. (1980) Going beyond interviewer observations of race in the National Health Interview Survey. In *Classification Issues in Measuring the Health Status of Minorities*. Hyattsville: National Center for Health Statistics.
Durrant, G. B., D'Arrigo, J. and Steele, F. (2011) Using paradata to predict best times of contact, conditioning on household and interviewer influences. *J. R. Statist. Soc.* A, **174**, 1029–1049.
Durrant, G. B., D'Arrigo, J. and Steele, F. (2012) Analysing interviewer call record data using a multilevel discrete time event history modelling approach. Submitted to *J. R. Statist. Soc.* A.
Durrant, G. B., Groves, R. M., Staetsky, L. and Steele, F. (2010) Effects of interviewer attitudes and behaviors on refusal in household surveys. *Publ. Opin. Q.*, **74**, 1–36.
Durrant, G. B. and Steele, F. (2009) Multilevel modelling of refusal and non-contact non-response in household surveys: evidence from six UK Government surveys. *J. R. Statist. Soc.* A, **172**, 361–381.
Faraway, J. J. (2005) *Linear Models with R*. Boca Raton: Chapman and Hall–CRC.
Feldman, J. J., Hyman, H. and Hart, C. W. (1951) A field study of interviewer effects on the quality of survey data. *Publ. Opin. Q.*, **15**, 734–761.
Fuller, W. (1987) A single explanatory variable. In *Measurement Error Models*, ch. 1. New York: Wiley.
Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Publ. Opin. Q.*, **70**, 646–675.
Groves, R. M. and Heeringa, S. G. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. R. Statist. Soc.* A, **169**, 439–457.
Groves, R. M., Mosher, W. D., Lepkowski, J. and Kirgis, N. G. (2009) Planning and development of the continuous National Survey of Family Growth. In *Vital Health Statistics*, vol. 1, no. 48. National Center for Health Statistics.
Hahn, R. A., Truman, B. T. and Barker, N. D. (1996) Identifying ancestry: the reliability of ancestral identification in the United States by self, proxy, interviewer, and funeral director. *Epidemiology*, **7**, 75–80.

Hill, M. E. (2002) Race of the interviewer and perception of skin color: evidence from the Multi-city Study of Urban Inequality. *Am. Sociol. Rev.*, **67**, 99–108.
Kott, P. S. (2006) Using calibration weighting to adjust for nonresponse and coverage errors. *Surv. Methodol.*, **32**, 133–142.
Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M. and Raghunathan, T. E. (2010) Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *J. R. Statist. Soc.* A, **173**, 389–407.
Landis, J. R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.
de Leeuw, E. and de Heer, W. (2002) Trends in household survey nonresponse: a longitudinal and international comparison. In *Survey Nonresponse* (eds R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), ch. 3. New York: Wiley.
Lepkowski, J. M., Mosher, W. D., Davis, K. E., Groves, R. M. and Van Hoewyk, J. (2010) The 2006–2010 National Survey of Family Growth: sample design and analysis of a continuous survey. In *Vital Health Statistics*, vol. 2, no. 150. National Center for Health Statistics.
Lessler, J. and Kalsbeek, W. (1992) Nonresponse: dealing with the problem. In *Nonsampling Errors in Surveys*, ch. 8. New York: Wiley-Interscience.
Little, R. J. A. (1986) Survey nonresponse adjustments for estimates of means. *Int. Statist. Rev.*, **54**, 139–157.
Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Hoboken: Wiley-Interscience.
Little, R. J. A. and Vartivarian, S. (2005) Does weighting for nonresponse increase the variance of survey means? *Surv. Methodol.*, **31**, 161–168.
Lynn, P. (2003) PEDAKSI: methodology for collecting data about survey non-respondents. *Qual. Quant.*, **37**, 239–261.
Matsuo, H., Billiet, J., Loosveldt, G. and Malnar, B. (2010) *Response-based Quality Assessment of ESS Round 4: Results for 30 Countries based on Contact Files*. Leuven: European Social Survey.
Mosteller, F. (1944) The reliability of interviewers' ratings. In *Gauging Public Opinion*, vol. 2, ch. 7, pp. 98–106. Princeton: Princeton University Press.
Nagelkerke, N. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691–692.
Pickering, K., Thomas, R. and Lynn, P. (2003) Testing the shadow sample approach for the English House Condition survey. *Report*. Office of the Deputy Prime Minister, London.
Särndal, C.-E. and Lundström, S. (2010) Design for estimation: identifying auxiliary vectors to reduce non-response bias. *Surv. Methodol.*, **36**, 131–144.
SAS Institute (2011) *SAS 9.2 Online Help and Documentation: the GLIMMIX Procedure (Default Output)*. Cary: SAS Institute.
Stähli, M. E. (2011) Examples and experiences from the Swiss interviewer training on observable data (neighborhood characteristics) for ESS 2010 (R5). *National Coordinators Meet. Mannheim, Mar. 31st–Apr. 1st*.
StataCorp (2009) *Stata Statistical Software: Release 11*. College Station: StataCorp.
Stefanski, L. A. and Carroll, R. J. (1985) Covariate measurement error in logistic regression. *Ann. Statist.*, **13**, 1335–1351.
Steiner, P. M., Cook, T. D. and Shadish, W. R. (2011) On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *J. Educ. Behav. Statist.*, **36**, 213–236.
Tolonen, H., Helakorpi, S., Talala, K., Helasoja, V., Martelin, T. and Prattala, R. (2006) 25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey. *Eur. J. Epidem.*, **21**, 409–415.
Valliant, R. (2004) The effect of multiple weighting steps on variance estimation. *J. Off. Statist.*, **20**, 1–18.
West, B. T. (2010) A practical technique for improving the accuracy of interviewer observations: evidence from the National Survey of Family Growth. *Technical Report 10-013*. Population Studies Center, Institute for Social Research, University of Michigan, Ann Arbor. (Available from `http://www.psc.isr.umich.edu/pubs/abs/6705`.)
West, B. T. and Little, R. J. A. (2011) Nonresponse adjustment of survey estimates based on auxiliary variables subject to error. Submitted to *Appl. Statist*.
Wolter, K. M. (2007) *Introduction to Variance Estimation*, 2nd edn. New York: Springer.
Zhang, D. and Lin, X. (2008) Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. *Random Effect and Latent Variable Model Selection* (ed. D. B. Dunson), ch. 2. New York: Springer.