Observational strategies associated with increased accuracy of interviewer observations: Evidence from the National Survey of Family Growth

Brady T. West[1], Frauke Kreuter[2]
[1] Michigan Program in Survey Methodology, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48104
[2] Joint Program in Survey Methodology, 1218 LeFrak Hall, College Park, MD, 20742

Abstract
Survey researchers are increasingly interested in obtaining as much auxiliary information on all sample units as possible, given trends of declining response rates in surveys of all formats worldwide. Surveys conducted using in-person (or "face-to-face") interviews therefore often request that interviewers collect observations on key features of selected households, given that interviewers are the eyes and ears of the survey organization out in the field. Unfortunately, recent studies have suggested that these observations are prone to error, and that the errors may decrease the effectiveness of nonresponse adjustments based in part on the observations. In addition, no studies have investigated the strategies being used by interviewers in the field when they are making these observations, or addressed the question of whether certain observational strategies improve the accuracy of the observations. This paper examines the associations of observational strategies used by field interviewers collecting face-to-face interviews from a large area probability sample with the accuracy of observations collected by those interviewers. A qualitative analysis shows that certain strategies are in fact associated with increased accuracy of the observations, and recommendations for future practice are presented.

Key Words: Interviewer Observations, Auxiliary Variables, Survey Data Collection

1. Introduction

Interviewers in "face-to-face" surveys are often charged with observing key features of selected households, given that interviewers are the eyes and ears of the survey organization out in the field. This is done in an effort to obtain as much auxiliary information on all sample units as possible, given trends of declining response rates in surveys of 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). 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) for reduction of both bias and variance in the estimates. Survey researchers will therefore request that interviewers attempt to collect observations on auxiliary variables having these optimal properties (in theory) for both respondents and nonrespondents.

The various conceptualizations of total survey error (TSE) that have been published over the years (see Groves and Lyberg, 2010, for a recent review) consistently acknowledge the problem of nonresponse bias that can arise in surveys. Errors in estimation (e.g., incorrect computation of the weights used for estimates, failure of analysts to correctly account for sample design features,

etc.), unfortunately, are less often acknowledged as a key part of TSE (see Deming, 1944, or Biemer, 2010, who refers to this problem as a type of data-processing error). 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, Ch. 8, p. 190; Stefanski and Carroll, 1985; West, 2010). The survey methodology literature has only recently begun to examine the non-negligible error properties of these observations (Campanelli et al., 1997, Chapter 4; Casas-Cordero, 2010; Groves et al., 2007; McCulloch et al., 2010; Pickering et al., 2003; Tipping and Sinibaldi, 2010; West, 2010), and no existing literature has assessed the observational strategies being used by field interviewers and whether different strategies in a given survey context lead to more or less accurate observations.

This research note presents an initial examination of observational strategies that are associated with the accuracy of a key interviewer judgment in the National Survey of Family Growth (NSFG). In the recently completed seventh cycle of the NSFG (June 2006 – June 2010; see Lepkowski et al., 2010, for design details), interviewers were tasked with judging whether a person between the ages of 15 and 44 who was randomly selected from an initial household screening interview was currently in a sexually active relationship with a member of the opposite sex. This judgment of perceived sexual activity was theoretically important due to its association with a number of key NSFG variables, along with the propensity of persons screened in the NSFG to respond to the survey request (West, 2010). These interviewer judgments could also be validated based on actual respondent reports of sexual activity collected in the main NSFG interview. NSFG interviewers in the last two quarters of data collection were asked to provide open-ended justifications for their sexual activity judgments, and the observational strategies indicated by these justifications are the main focus of this research. 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 address the following research questions:

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

2. Data

In the last two quarters of data collection for the NSFG, 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 Appendix Figure A3). The interviewers were asked to provide justifications right after the judgments were made. However, they were not prompted for specific justifications or limited in any way. This resulted in the collection of 3,992 open-ended justifications of widely varying lengths from the 45 interviewers during these two quarters of data collection[1]. Two examples of these 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.)
2. *"He has a tattoo, 'Carol', over his heart."* (A justification for a judgment of currently sexually active.)

[1] Analyses in this study were conducted at the interviewer level. Ongoing work is applying multilevel modeling techniques to examine respondent- and interviewer-level predictors of the accuracy in interviewer observations collected across all 16 quarters of the seventh cycle of the NSFG.

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, alone, etc.)
- Relationship status (mention of spouse, partner, etc.)
- Age
- Household characteristics (presence of children, cultural icons, cleanliness, etc.)
- Appearance (references to physical appearance, ethnicity, or pregnancy)
- Neighborhood characteristics
- Shyness
- Guess (indication of a gut feeling, or not being sure)
- Incorrect (mention that an incorrect observation was entered in hindsight)
- Conservative (indication of conservative or strict household / parents)
- Health (reference to the health or physical disability of the person)
- Personality (reference to the person's personality or general demeanor)
- Occupation (reference to the person's occupation or student status)

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[2]. Discrepancies in coding were detected using the COMPARE procedure 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.

To compute dependent variables for each interviewer, the total number of judgments of sexual activity across these two NSFG quarters, 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 also recorded 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 based on survey reports who were judged to be sexually active), and the *false negative rate* (i.e., the proportion of sexually active respondents based on survey reports 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.

The University of Michigan Institutional Review Board (IRB) approved the design of this study.

3. Methods

We used 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

[2] We are indebted to Ziming Liao from the University of Michigan Undergraduate Research Opportunity Program (UROP) for his contributions to this work.

interviewers were first standardized. An agglomerative hierarchical clustering approach was then applied (Everitt et al., 2011, Chapter 4), 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).

Each of the three interviewer-level dependent variables (gross difference rate, false positive rate, and false negative rate) were regressed on a categorical variable indicating cluster membership using the LOGISTIC procedure in the SAS software (Version 9.2, SAS Institute, Cary, NC). Specifically, models of the following form were fitted (using gross difference rate as an example), omitting an indicator variable for one of the identified clusters (as the reference category):

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{c=1}^{C-1} \beta_c I_i\left[cluster = c\right]$$

where

p_i = probability of incorrect judgment for IWER i

$$I_i\left[cluster = c\right] = \begin{cases} 1 \text{ if IWER } i \text{ in cluster } c \\ 0 \text{ otherwise} \end{cases} \qquad (1)$$

4. Results

We first consider a descriptive summary of the variables computed for the 45 interviewers. Descriptive statistics for the interviewer-specific percentages and mean word counts are shown in Table 1 below.

Table 1: Descriptive statistics for variables used in the cluster analysis: interviewer-level justification tendencies (in descending order by mean percentages of justifications) and mean word counts

	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
Household Characteristics	23.75	13.18	0.00	62.50
Guess	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.

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 (see Table 1). Interviewer justifications for sexual activity judgments most often referred to the perceived relationship status of selected respondents. All interviewers used this justification for at least some of their observations and one of them for as many as 75% of the justifications made. Roughly 1% of the justifications (about 40 justifications) indicated that an incorrect judgment was entered in hindsight.

Table 2 below presents descriptive statistics for the three dependent variables computed for each interviewer.

Table 2: Descriptive statistics for 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, which could not be computed for two interviewers who did not have any sexually inactive respondents.

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.

Given the error rates listed in Table 2, it is particularly interesting to see how the justifications given by the interviewers relate to judgment accuracy. An initial cluster analysis provided evidence of two interviewers that could be considered outliers (see Appendix Figure A1), with one interviewer citing neighborhood features in 55.41% of justifications (the next highest percentage being 17.12%), and another interviewer citing health reasons for 43.14% of justifications (the next highest percentage being 27.50%). After dropping these two interviewers, a second cluster analysis was performed that presented evidence of four clusters of interviewers based on scaled distances between the clusters (see Appendix Figure A2); that is, there were in fact distinct groups of interviewers in terms of justification tendencies. Descriptive statistics on the 14 variables for each cluster are shown in Table 3 below.

Table 3: Descriptive statistics for interviewer-level justification tendencies and mean word counts within four distinct clusters of interviewers

	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), $p < 0.001$
Household Characteristics	**28.60 (13.25)**	25.11 (9.33)	13.04 (10.81)	26.23 (14.42)	9.0 (3), $p = 0.029$
Guess	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). Significance tests use a non-parametric independent samples Kruskal-Wallis test with all pairwise comparisons.

The results in Table 3, with the largest cluster means for each justification indicator boldfaced in the case of significant differences in distributions across the four clusters, suggest that the first cluster of interviewers is largely defined by a tendency to notice living arrangement and housing characteristics. The second cluster is largely defined by references to appearance and personality, and a relatively large word count. 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.

Table 4 presents estimates of the coefficients in each of the three logistic regression models for the three error rates. First, in the case of the gross difference rates, a Wald chi-square test of the null hypothesis that all three regression parameters in the logistic regression model (indicating clusters 1, 2, and 3, with cluster 4 as the reference category) were equal to zero suggested a rejection of this null hypothesis, indicating that cluster membership was a strongly significant predictor of gross difference rates (Wald $\chi^2(3) = 23.53$, $p < 0.001$). After applying a conservative Bonferroni adjustment given all six possible pairwise comparisons of the gross difference rates between the four clusters (implying a critical level of $0.05 / 6 = 0.0083$ for the significance of a given comparison), cluster 1 (largely defined by interviewers tending to use justifications based on living arrangement and housing characteristics) had a significantly higher gross difference rate than clusters 2 and 3, and the remaining clusters had comparable gross difference rates. In the case of the false positive rates, cluster membership was once again a significant predictor (Wald $\chi^2(3) = 23.63$, $p < 0.001$), with cluster 4 having a significantly higher false positive rate than clusters 1, 2, and 3 after adjusting for the multiple comparisons. Finally, in terms of the false negative rates, cluster membership was once again a significant predictor (Wald $\chi^2(3) = 71.15$, $p < 0.001$), with clusters 2, 3, and 4 all having significantly lower false negative rates than cluster 1, with cluster 4 also having a significantly lower false negative rate than cluster 2.

Table 4: Parameter estimates in logistic regression models indicating relationships of interviewer 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**	
	Estimate (SE)	Test Statistic	Estimate (SE)	Test Statistic	Estimate (SE)	Test Statistic
Cluster Membership		$\chi^2_3 = 23.53$, $p < 0.001$		$\chi^2_3 = 23.63$, $p < 0.001$		$\chi^2_3 = 71.15$, $p < 0.001$
Intercept	-1.11 (0.06)	$\chi^2_1 = 355.8$,	-0.35 (0.11)	$\chi^2_1 = 10.82$,	-1.41 (0.07)	$\chi^2_1 = 369.9$,

		$p < 0.001$		$p = 0.001$		$p < 0.001$
Cluster 1	--	--	--	--	--	--
Cluster 2	-0.33 (0.12)	$\chi^2_1 = 7.15$, $p = 0.008$	0.49 (0.21)	$\chi^2_1 = 5.63$, $p = 0.018$	-0.95 (0.19)	$\chi^2_1 = 24.98$, $p < 0.001$
Cluster 3	-0.49 (0.11)	$\chi^2_1 = 18.12$, $p < 0.001$	0.41 (0.19)	$\chi^2_1 = 4.75$, $p = 0.029$	-1.17 (0.18)	$\chi^2_1 = 43.69$, $p < 0.001$
Cluster 4	-0.47 (0.19)	$\chi^2_1 = 5.94$, $p = 0.015$	1.71 (0.39)	$\chi^2_1 = 19.33$, $p < 0.001$	-3.70 (1.01)	$\chi^2_1 = 13.56$, $p < 0.001$

-- indicates reference category.

When considering these findings alongside the results in Table 3, we find consistent evidence of the different clusters of interviewers varying in terms of their error rates; that is, observational strategies influenced the error properties of the sexual activity judgments in a significant manner. These results suggest that:

- focusing primarily on relationship status and gut feelings / guessing (cluster 3) will result in the highest accuracy (an overall gross difference rate of 0.168) and more variable errors (i.e., a lower false positive rate and a higher false negative rate relative to cluster 4);
- judging based primarily on age (cluster 4) will result in relatively high accuracy (an overall gross difference rate of 0.171) and systematic false positives;
- focusing primarily on appearance, personality, and other external features (cluster 2) will result in slightly lower accuracy and more variable errors;
- focusing on living arrangement and household features (cluster 1) is detrimental in terms of accuracy and results in systematic false negatives.

5. Discussion

Given existing theory (Lessler and Kalsbeek, 1992, Ch. 8, p. 190; Stefanski and Carroll, 1985) and previous simulation work (West, 2010), the use of error-prone interviewer observations to adjust survey estimates for nonresponse has the potential to reduce the quality of the estimates. Techniques for reducing the non-negligible levels of error in these observations (see West, 2010) therefore require more research focus from survey methodologists. This research note has demonstrated that the collection and analysis of open-ended justifications for the observations and judgments that field interviewers are making while conducting face-to-face surveys is feasible in practice. The analyses presented in this study provide interesting insights into the observational methods used in the field by NSFG interviewers who make more accurate judgments regarding the perceived current sexual activity of screened persons.

With respect to the first research question posed, we did find 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 provided by NSFG staff, different interviewers did in fact use different observational strategies in the field when recording these types of judgments; this finding certainly needs to be replicated in other survey contexts to further understand this phenomenon. With respect to the second research question posed, the four clusters of interviewers identified based on their justification strategies were found to vary significantly in terms of the error properties of this specific interviewer observation. This finding suggests that variance in error rates on these types of observations may in fact be a function of the observational strategies being used in the field, and this finding also needs to be replicated in other survey contexts.

Although all of the NSFG interviewers were asked to record these judgments based on their initial impressions and best guesses, which is suggested in the social psychology literature to be beneficial for the accuracy of these types of brief observations (Ambady et al., 1999; Patterson and Stockbridge, 1998), the more accurate NSFG interviewers tended to focus on relationship status and age (as determined from the screening interview[3]), which are certainly key correlates of current sexual activity, while minimizing attention to living arrangement. Guessing based on gut feelings and brief impressions was also found to not be entirely detrimental to accuracy rates, as suggested by literature in social psychology (Ambady et al., 1999); however, coding of the justifications revealed that some guesses were based on *not having seen or met the selected respondent* (this was not explicitly coded for analysis, but was sometimes mentioned for guesses). This is a serious problem with the technique of making doorstep observations when screening interviews are conducted with *informants* rather than the eventual respondents, and future research in this area might contrast the error properties of judgments based on proxy informants with those of judgments based on actual respondents. This issue aside, the results in this research note could be used to guide future interviewer training sessions, and the analytic techniques used may serve to inform effective observational techniques in other survey contexts.

References

Ambady, Nalini, Hallahan, Mark and Brett Conner. (1999). Accuracy of Judgments of Sexual Orientation from Thin Slices of Behavior. *Journal of Personality and Social Psychology*, 77(3), 538-547.

Baruch, Yehuda and Brooks C. Holtom. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, 61(8), 1139-1160.

Beaumont, Jean-Francois. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31(2), 227-231.

Bethlehem, Jelke G. (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, Paul P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, 74(5), 817-848.

Biener, Lois, Garrett, Catherine A., Gilpin, Elizabeth A., Roman, Anthony M., and Douglas B. Currivan. (2004). Consequences of Declining Survey Response Rates for Smoking Prevalence Estimates. *American Journal of Preventative Medicine*, 27(3), 254-257.

Campanelli, Pam, Sturgis, Patrick, and Susan Purdon. (1997). *Can you hear me knocking: An investigation into the impact of interviewers on survey response rates*. London: SCPR.

Casas-Cordero, Carolina. (2010). Assessing the quality of interviewer observations of neighborhood characteristics. *Paper presented at the 2010 International Total Survey Error Workshop*, Stowe, Vermont, June 14, 2010.

[3] This finding suggests that the availability of demographic features on a sampling frame for persons in a target population of interest may eliminate the need for interviewer observations, if the demographic information is strongly associated with key survey variables.

Cull, William L., O'Connor, Karen G., Sharp, Sanford, and Suk-fong S. Tang. (2005). Response rates and response bias for 50 surveys of pediatricians. *Health Services Research*, 40(1), 213-226.

Curtin, Richard, Presser, Stanley, and Eleanor Singer. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, 69, 87–98.

de Leeuw, Edith, and Wim de Heer. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Chapter 3 in Groves, R.M. et al., *Survey Nonresponse*. Wiley.

Deming, W. Edwards. (1944). On Errors in Surveys. *American Sociological Review*, 9, 359-369.

Everitt, Brian S., Landau, Sabine, Leese, Morven and Daniel Stahl. (2011). *Cluster Analysis, 5th Edition*. Wiley Series in Probability and Statistics.

Groves, Robert M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5), 646-675.

Groves, Robert M. and Lars Lyberg. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5), 849-879.

Groves, Robert M., Wagner, James, and Emilia Peytcheva. (2007). 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.

Lepkowski, James M., Mosher, William D., Davis, Karen E., Groves, Robert M., and John Van Hoewyk. (2010). 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, 2(150), June 2010.

Lessler, Judith and William Kalsbeek. (1992). Nonresponse: Dealing with the Problem. Chapter 8 in *Nonsampling Errors in Surveys*. Wiley-Interscience.

Little, Roderick J., and Sonya Vartivarian. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31(2), 161-168.

McCulloch, Susan K., Kreuter, Frauke, and S. Calvano. (2010). 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.*

Patterson, Miles L. and Erica Stockbridge. (1998). Effects of cognitive demand and judgment strategy on person perception accuracy. *Journal of Nonverbal Behavior*, 22(4), 253-263.

Pickering, Kevin, Thomas, Roger, and Peter Lynn. (2003). 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.*

Punj, Girish and David W. Stewart. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*, 20(2), 134-148.

Sinibaldi, Jennifer. (2010). Measurement Error in Objective and Subjective Interviewer Observations. *Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research, Chicago, IL, 5/14/2010.*

Stefanski, Leonard A., and Raymond J. Carroll. (1985). Covariate Measurement Error in Logistic Regression. *The Annals of Statistics*, 13(4), 1335-1351.

Tipping, Sarah and Jennifer Sinibaldi. (2010). 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.

Tolonen, Hanna, Helakorpi, Satu, Talala, Kirsi, Helasoja, Ville, Martelin, Tuija, and Ritva Prattala. (2006). 25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey. *European Journal of Epidemiology*, 21, 409-415.

Ward, Joe H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

West, Brady T. (2010). An Examination of the Quality and Utility of Interviewer Estimates of Household Characteristics in the National Survey of Family Growth. NSFG Paper No. 10-009. May 2010. *Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research, Chicago, IL, 5/14/2010.*

Appendix

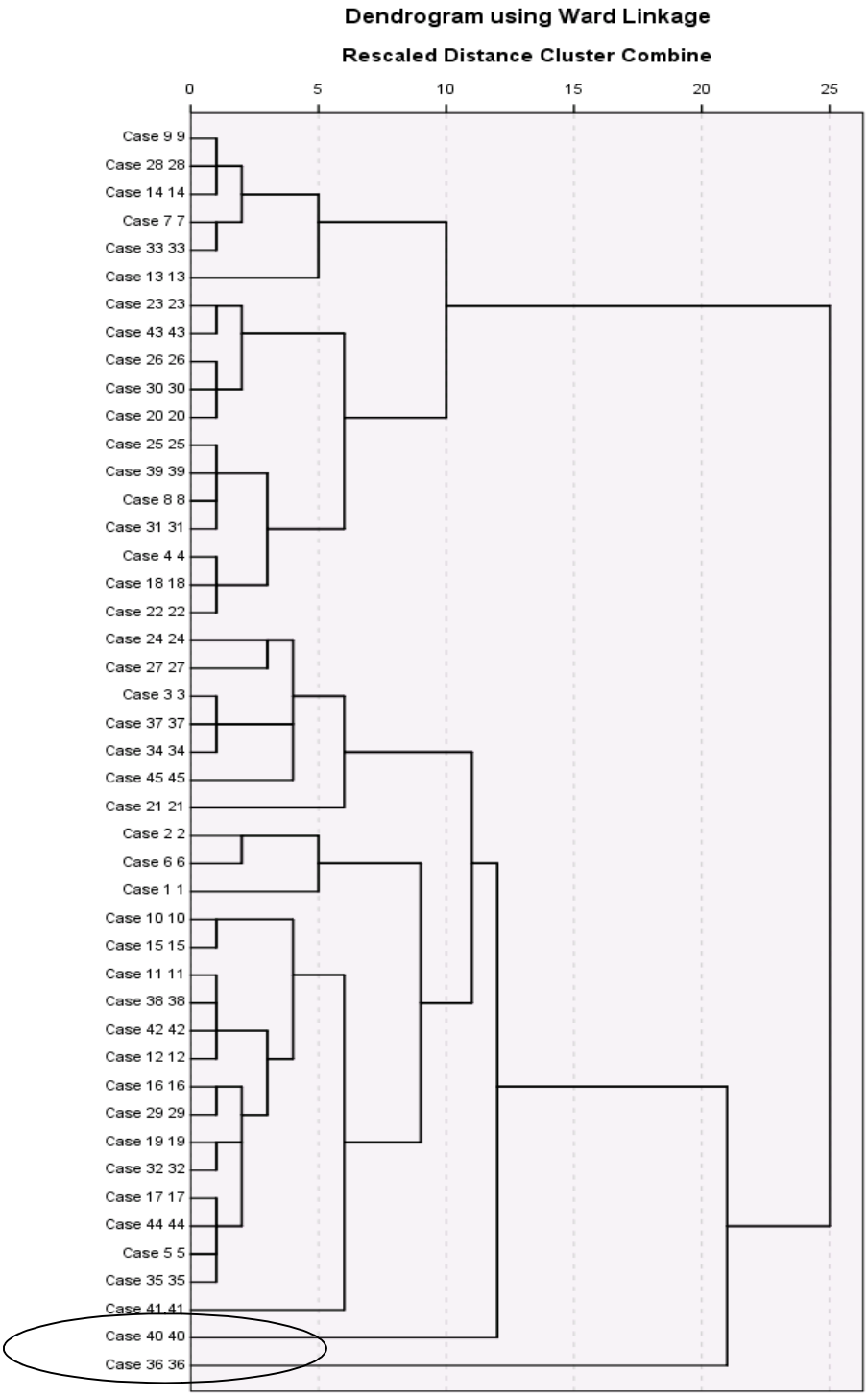

Figure A1: Dendrogram showing results of initial hierarchical agglomerative cluster analysis, with evidence of two outliers (Interviewers 36 and 40)

Figure A2: Dendrogram showing results of second cluster analysis (excluding the two outliers), with evidence of four distinct groups of interviewers (based on rescaled cluster distances greater than 10)

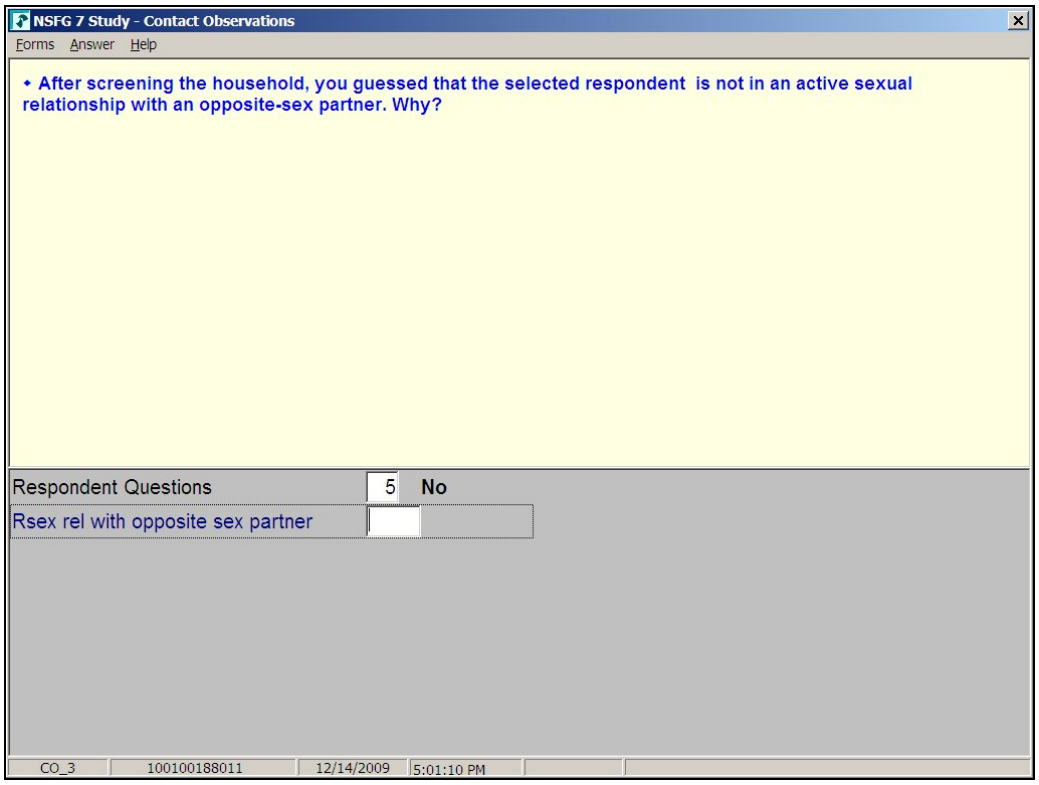

Figure A3: Screenshot of CAPI application screen where NSFG interviewers could enter open-ended justifications for their sexual activity judgments. The justification was typed into the box labeled "Rsex rel with opposite sex partner," which was converted to a full text box when interviewers started typing