EVALUATION OF THE INTERVIEWER QUALITY CONTROL PROCEDURE FOR THE POST-ENUMERATION SURVEY*

Lynne Stokes, University of Texas, and Patty Jones, U.S. Bureau of the Census
Lynne Stokes, MSIS Dept., UT, Austin, TX 78712

KEY WORDS: Curbstone, Reinterview, Logistic Regression

1. Introduction

In any large survey in which the data are collected by interviewers, there is a danger that some of them may fabricate data. The U.S. Bureau of the Census refers to this practice as curbstoning, and the interviewers who do it as curbstoners. A substantial amount of resources were committed to detecting and eliminating curbstoning in the Post-Enumeration Survey (PES) in the Census test sites in 1986 and 1988. The purpose of this paper is to describe one method for evaluating these detection programs, and to discuss its results from the Los Angeles Test of Adjustment Related Operations and the Saint Louis Dress Rehearsal. We also discuss some of what was learned about PES curbstoners in the process.

Section 2 below explains why minimizing curbstoning is of particular importance for the PES. Section 3 briefly describes the curbstoner detection procedure that was used for interviewers in the PES. The method we developed for evaluation of the procedure and the results of that evaluation are discussed in Section 4. A discussion of the results, including implications for a redesign of the reinterview sampling procedure, follow in Section 5.

2. The Impact of Curbstoners on PES Data

The few studies of curbstoners which have been reported agree that they make up a small minority of interviewers. A recent study of data from some of the Census Bureau's current surveys reports that between 2 and 5% of interviewers cheat in some way in data collection (Biemer and Stokes 1989). Most of these infractions involved fabrication of interviews, but certainly not of their entire assignments. From the data of that study, we guess that between 1/2% and 1 1/2% of the interviews themselves are fabricated in those surveys. This fabrication rate may seem reassuringly small. However, there are several circumstances which suggest that the potential impact on the PES may be greater than these small numbers would lead one to expect.

First, the same study of Census Bureau curbstoners showed that inexperienced interviewers are far more likely than experienced ones to fabricate responses. Since PES interviewers will be temporary employees, they are likely to be inexperienced. So the fabrication rate for the PES may be higher than is generally experienced in other Census Bureau surveys. The second caution is that fabricated units invariably introduce a one-sided bias into the dual system estimator, which is used for estimating the undercount from PES data. Any undetected fabricated unit in the PES would clearly not match to the Census and would therefore falsely inflate the number of nonmatches and, in turn, the undercount estimate. But even worse is the fact that this bias is likely to be differential, meaning that it will be larger for some poststrata than for others. It is generally believed that interviewers are most likely to curbstone in hard-to-enumerate areas, where the nonmatch rate is likely to be high. If this is the case (in Section 5 we discuss some evidence from this study suggesting it is), then those strata having the highest undercounts will have the largest bias, while those with the smallest undercounts will have the least bias. This will falsely exaggerate the undercount variability.

3. The Reinterview Procedures

The procedures used for detecting curbstoning in all five test sites (Columbia, MO; North Dakota; and Eastern Washington, in addition to the two already mentioned) were similar. Reinterviews with a subset of the PES sample were performed. That subset was selected by a stratified systematic design, where the strata were interviewer work units. A work unit consisted of all interviews and vacant housing units completed by an interviewer within the same geographic area. Work units usually contained one or two days of an interviewer's work. An average of about 1/3 of the households were selected for reinterview, with a higher rate used for small work units and a lower rate for large work units. A telephone call or personal visit was made to each sample case in an attempt to determine if the right household was interviewed and whether all and only the true household members were included on the PES roster of names. If any sample case failed the roster check, then all cases in that work unit were reinterviewed. If serious errors were found, a portion of some previous work units were also reinterviewed.

This intense procedure detected only three suspected curbstoners. All three were found in the two urban areas, Los Angeles and Saint Louis. Further, one of the two cases in Los Angeles appeared to result from a failure to properly follow probing procedures, rather than from blatant curbstoning (Corby 1987).

4. A Method for Evaluating the Detection Procedure

Despite the careful procedure just described, it is possible that some PES curbstoners remained undetected. Even though the work of all interviewers was sampled, an existing fabricated unit might not have been selected or identified even if reinterviewed. One way of gaining information about if (or perhaps more reasonably how often) this occurred is to compare the nonmatch rates for the interviewers' assignments. The Biemer and Stokes study (1989) showed that inexperienced interviewers who were detected curbstoning fabricated an average of about 30% of the units in their assignment, while the rate for their experienced counterparts was about half that. Such a high level of curbstoning should be reflected in a high total household nonmatch rate for these interviewers.

There are at least two problems with simply ranking interviewers on the basis of their total household nonmatch rates. First, interviewer assignments are not equally difficult. We know that there are some types of households that are harder to enumerate than others. Interviewers whose assignments contain households of this type will naturally have higher nonmatch rates than interviewers with easier assignments, even if they follow procedures carefully. The second problem is that interviewer assignment sizes varied considerably in both Los Angeles and Saint Louis. The observed nonmatch rates are poor estimates of true rates for interviewers whose assignments contained a small number of households.

To handle the first problem, we built a model to predict the probability of a nonmatch from available characteristics of each household. We restricted attention to nonmover households only in this analysis. The model was applied to each nonmover household in an interviewer's assignment, and the resulting predicted probabilities were averaged to

obtain an expected nonmatch rate for the interviewer. The idea, then, was to compare each interviewer's expected and observed rates, and only when the latter substantially exceeded the former would there be serious concern regarding the interviewer.

We built separate logistic regression models for predicting household nonmatch rate from the data for Los Angeles and Saint Louis. (We excluded the data of the three suspected curbstoners in this step.) A household nonmatch was defined to be a household in which no members matched to the Census. The resulting models for the two sites were similar, but not identical. A well-fitting model for nonmatch rate in Saint Louis was obtained using the predictors rental status (rental/nonrental), race (black/nonblack), and household size. The analysis of variance table for this model is shown in Table 1. The table shows that the household nonmatch rate is higher for renters and blacks, and decreases with increasing household size. There was no evidence of interaction among any pair of the variables. No other variables were found to improve the prediction beyond what these provided. Others tested included proxy respondent status, age and sex of respondent, and building type (single family versus multiple unit dwellings).

Table 1. ANOVA for Model for Household Nonmatch Rates in Saint Louis

Source	Estimate	s.e.	Chi-Square	p-value
Intercept	-1.97	.14	197.61	.0001
Household Size	-.38	.06	36.55	.0001
$(\text{HH Size})^2$	.02	.01	13.40	.0003
Race (Black = 1)	.51	.12	18.66	.0001
Rental Status (Rent = 1)	.41	.11	13.53	.0002

It was striking that the only two useful predictors (besides household size) of matching status were exactly those block characteristics selected as stratifying variables for the PES in Saint Louis; i.e., rental status and race (Thompson 1987). Our investigation therefore suggests that the stratification design should have been very effective in improving estimates of nonmatch rates. The PES stratification design included only the three strata black renter, black owner, and nonblack, however, while we found evidence that the rental status of nonblacks was predictive of nonmatch rate as well.

For the Los Angeles data, we found only race and rental status of households to be predictive of their match status, and not their size. There were few blacks in the Los Angeles sample, however, and the useful dichotomy of the race variable for prediction purposes was other/not other. Apparently, the "other" category of race served as a proxy for Hispanic ethnicity. It was better, in fact, than the ethnicity question for prediction of match status. Table 2 shows the analysis of variance table for the model. It shows that a household was more likely to be a nonmatch if the housing unit was rented or if the race classification was Other.

Table 2. ANOVA for Model for Household Nonmatch Rates in Los Angeles

Source	Estimate	s.e.	Chi-Square	p-value
Intercept	-2.91	.16	329.84	.0001
Race (Other = 1)	.32	.14	5.11	.0239
Rental Status (Rent = 1)	1.06	.17	38.30	.0001

The difference in the effectiveness of household size as a predictor of nonmatch status in Los Angeles and Saint Louis is notable. Because of the way a nonmatch is defined (as a household in which *no* person matches), it is reasonable to expect that the nonmatch rate should be decreasing with household size. This relationship was very clear and strong in the Saint Louis data, but was missing entirely in the Los Angeles data. Unfortunately, we have no explanation for this difference.

Our next step was to use the models for determining an expected nonmatch rate for each interviewer. The probability of a nonmatch was predicted for each household and these predictions averaged over the nonmover households in each interviewer's assignment. The results of this step confirmed that interviewer assignments varied widely with respect to their expected nonmatch rate. Table 3 shows some characteristics of these rates for the 88 interviewers in Saint Louis and the 43 in Los Angeles who had assignments of at least ten nonmover households. Note that the expected household nonmatch rates ranged from a low of .06 in both sites to a high of about three times that (.15 in Saint Louis and .16 in Los Angeles).

Table 3. Expected Household Nonmatch Rates for Interviewers

Site	# of Interviewers	Mean	St.Dev.	Minimum	Maximum
L.A.	43	.11	.03	.06	.16
St. Louis	88	.11	.02	.06	.15

Finally, we compared each interviewer's expected and observed nonmatch rates. Comparing the absolute difference of these rates would not be appropriate because of the large variance of the observed rates for those interviewers with small assignments. Therefore, we computed for each interviewer a score of the form

$$z_i = (\bar{p}_i - e_i/n_i)/[\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})]^{1/2}/n_i;$$

where n_i is the number of households in the ith interviewer's assignment, e_i is the number of nonmatches among them, p_{ij} is the predicted probability of a nonmatch for the jth household in that assignment, $\bar{p}_i = \sum_{j=1}^{n_i} p_{ij}/n_i$. If the model holds and if n_i is moderately large, we should expect z_i to be approximately normally distributed. Thus we might investigate as possibly discrepant an interviewer for whom $|z_i|$ is large, say greater than 3. We hoped to find that only the suspected curbstoners already identified had discrepantly high (negative z_i) nonmatch rates.

Table 4 displays the z values, the expected and observed nonmatch rates, and the sample sizes for the ten most discrepant (both high and low nonmatch rates) interviewers. The interviewers marked with an asterisk are the suspected curbstoners. It is clear from the table that all three suspected curbstoners were highly discrepant with respect to total household nonmatch rate. However, one other interviewer in Saint Louis also had a suspiciously high rate; it is higher, in fact, than that of the suspected curbstoner. An investigation of the records from the reinterviewing activities showed no indication that this interviewer was ever suspected of curbstoning. We could also uncover no other reason (not captured by our model) why this assignment should have had an especially high nonmatch rate, such as its being located in a university or migrant area. We are left to conclude that this may have been an undetected curbstoner. In general, however, we feel that our evaluation showed that the detection method performed well in identifying at least the most damaging curbstoners, which are those who fabricate large numbers of units.

Table 4. Ten Most Discrepant Interviewers in Each Site

Saint Louis				Los Angeles			
n_i	$\bar{p}_i$	e_i/n_i	z_i	n_i	$\bar{p}_i$	e_i/n_i	z_i
High Nonmatch Rates							
36	.12	.44	-5.9	63	.16	.41	-5.4*
61	.12	.30	-4.2*	82	.12	.27	-4.1*
16	.09	.31	-3.0	31	.12	.29	-3.0
114	.08	.15	-2.6	106	.12	.19	-2.1
58	.13	.22	-2.2	22	.11	.23	-1.7
Low Nonmatch Rates							
38	.07	.00	1.7	63	.14	.05	2.1
56	.08	.02	1.7	80	.10	.03	2.2
65	.10	.03	1.9	59	.10	.02	2.2
32	.12	.00	2.1	50	.14	.02	2.4
75	.09	.01	2.4	81	.14	.02	3.0

It is also satisfying to note that no discrepantly high z values, indicating far better than expected matching performance, occurred in either site. This would be expected if the procedure were performing adequately. There should be no special method for an interviewer to get fewer nonmatches than would be obtained by simply following procedures as written, which all well-trained interviewers should do.

5. Strategies for Detecting Curbstoners

Though testing of PES procedures, including the reinterviewing process, occurred in five sites, the only sites in which suspected curbstoners were detected were Saint Louis and Los Angeles. It has long been believed that curbstoning results from pressures on interviewers to obtain interviews in difficult-to-enumerate areas. Since urban areas are notorious for their difficulty, the presence of detected curbstoners in the two urban sites and their absence from the nonurban ones are supportive of this belief.

But this study suggests that even within an urban site, curbstoning occurs most frequently in difficult areas, or at least by interviewers whose assignments have high expected nonmatch rates. The support for this statement is from Tables 3 and 4, which show the expected nonmatch rates, as computed from our models, for the discrepant interviewers and for all interviewers working in each site. All three interviewers detected as possible curbstoners during reinterviewing were well above average with respect to their expected nonmatch rate (with rates of .12, .12, and .15), as was the additional possible curbstoner (with a rate of .12) identified by our evaluation process in Saint Louis.

The PES reinterview procedure has three goals. The first is to detect interviewers who fabricate responses as soon as possible, and to replace those responses with valid data. The second is to deter interviewer cheating by the knowledge that a checking procedure is in use. The third is to collect data that may be used to estimate fabrication rates, so that these rates can be used in evaluation of the quality of the undercount estimates. The best reinterview sample designs for meeting these goals are sometimes in conflict. The first goal would demand that the largest sampling rates occur in areas where curbstoning is most frequent, while the third goal would require that the largest samples were in areas where fabrication rates are small and therefore more difficult to estimate. The best design for meeting the middle goal is not so clear, but it seems reasonable to believe that assuring that each interviewer's work is sampled (and that the interviewing staff is aware of this policy) would be desirable. The design used in the 1986 and 1988 PES's took a middle course, which was to use an approximately equal selection rate for the reinterview sample for all interviewers. Our recommendation is that the design be constructed to facilitate the first and second goals; i.e., those of identifying and eliminating as many fabricated units as possible, while at the same time ensuring that some of each interviewer's work is subject to reinterview. One practical way of approaching this would be to retain interviewer work units as strata, but use varying sampling rates. The smallest rates could be used in non-urban areas and low risk urban blocks, while the highest rates could be reserved for the remaining urban blocks. Information about race and rental status at the block level is available prior to sampling, so that such a design could be easily implemented in 1990, especially with the planned automated sample selection procedures.

References

Biemer, P.P. and Stokes, S.L. (1989) "The Optimal Design of Quality Control Samples to Detect Interviewer Cheating," Journal of Official Statistics 5, 23-39.

Corby, Carol (1987) U.S. Bureau of the Census Memorandum to the Record, "Documentation of the Quality Control Operations for the 1986 Post Enumeration Survey," April 19, 1987.

Thompson, John (1987) U.S. Bureau of the Census, STSD 1988 Dress Rehearsal Memorandum Series #V-4,"Documentation of the Sample Selection for the 1988 Dress Rehearsal Post Enumeration Survey," December 7, 1987.

*The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.