Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 433–450

Using Statistical Models for Sample Design of a Reinterview Program

Jianzhu Li[1], *J. Michael Brick*[1,2], *Bac Tran*[3], *and Phyllis Singer*[3]

The U.S. Census Bureau relies on reinterview programs as the primary method to evaluate field work and monitor the work of the interviewers. One purpose of the reinterviews is to identify falsification. Since falsification is a rare occurrence, reinterview programs generally identify very few falsified cases even when the reinterview sample is reasonably large. This study examines methods for designing a reinterview sample with the goal of identifying more falsified cases. With the Current Population Survey (CPS) as an example, we explore data that could be used for reinterview sampling beyond that currently used in the CPS program. We fit a logistic regression model to predict the likelihood of falsification with the data from original interviews, and use the predicted probabilities to construct alternative reinterview sampling designs. The alternative designs are compared to the current sampling method through cross validation and simulation methods.

Key words: Quality control; falsification; curb stone; rare event.

1. Introduction

Interviewer falsification of survey data has been an important concern and area of study at survey organizations for many decades because of its potential effect on data quality. Interviewer falsification is defined as the knowing departure from the designed interviewing procedures or guidelines to avoid interviewing, classifying, and/or listing units (Schreiner et al. 1988). For example, interviewers may fabricate all or part of an interview, record a refusal case as ineligible for the sample, report a fictitious contact attempt, or accept proxy information from a person who is not eligible to respond. These activities are sometimes called "cheating" or "curb stoning."

Many surveys include a quality assurance program to help control the quality of the data collection process, and reinterviewing a subsample of the original sample is often the primary method used to evaluate field work or monitor the work of the

[1] Westat, 1600 Research Boulevard, Rockville, MD 20850 U.S.A. Emails: JaneLi@westat.com and mikebrick@westat.com
[2] Joint Program in Survey Methodology, University of Maryland, 1218 Lefrak Hall, College Park, MD 20742 U.S.A. Email: mikebrick@westat.com
[3] U.S. Census Bureau, 4700 Silver Hill Road, Washington, DC 20233-1912, U.S.A. Emails: bac.tran@census.gov and phyllis.singer@census.gov
Disclaimer: This article is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau or Westat.

interviewers. Reinterview programs may focus on identifying falsified interviews, but budget constraints and the low frequency of falsification result in reinterview samples that are often too small and inefficient to identify falsified survey data with a high probability.

In one of the earliest studies, Crespi (1945) conjectured that factors related to either the questionnaire or the administration of the survey may cause interviewers to falsify data. He suggested reducing falsification by careful and intelligent survey design and administration, along with a verification method. Although such approaches to survey design and administration have been addressed in many publications (Bennett 1948a and 1948b; Sheatsley 1951; Boyd and Westfall 1955; Evans 1961), these provide scant guidance on implementing verification methods.

Over the past three decades, researchers from the U.S. Census Bureau and other survey organizations have been actively exploring this area. The most commonly studied method for detecting falsified cases is reinterview. Other methods such as reviewing outliers and modeling the probability of falsification have also been explored. In 1982, the U.S. Census Bureau initiated an Interviewer Falsification Study to collect information on the confirmed or suspected cases of falsification by interviewers from its demographic surveys. Shreiner et al. (1988) analyzed the results from the first five years of the study and found: (1) the two most common falsification techniques are the complete fabrication of an interview and the misclassification of occupied housing units as vacant; (2) falsification rates are higher for new interviewers; and (3) the intra-interviewer correlation of falsification is very low (contrary to one of the assumptions Schäfer et al. 2004 made in their work on falsification). Some of these results are tenuous because the study only had 205 confirmed cases of falsification. In another study, Waite (1993; 1997) found that experienced interviewers were less likely to falsify, and falsified fewer of their cases.

Biemer and Stokes (1989) used the Interviewer Falsification Study to design a quality control sample based on the assumption that falsification is a random event characterized by parameters depending on the interviewer. They proposed designing reinterview samples based on this process so as to maximize the probability of detecting falsification for a fixed cost. A practical problem with their approach is that the parameters of the model are not known in advance, and prior information is difficult to obtain.

The focused reinterview method to select a reinterview sample was introduced in the 1990s (Hood and Bushery 1997; Krejsa et al. 1999; Bushery et al. 1999). The focused reinterview uses control charts to identify outcomes of interviewers that stand out for particular variables. The characteristics which were found to identify atypical interviewers include rate of ineligible units, rate of short interviews, rate of no telephone number, and rate of completed interviews. It was concluded that the focused reinterview appeared to detect falsification at a higher rate than the random reinterview.

Other efforts to detect falsification have suggested a systematic review of response data and metadata (e.g., Murphy et al. 2004). Swanson et al. (2003) and Schäfer et al. (2004) proposed using Benford's Law and other methods based on variability in reported continuous items to improve the detection of falsified cases; these approaches can only identify falsification in cases that are reported as being completed interviews. Schräepler and Wagner (2005) studied the potential effect that falsification has on estimates from a survey.

While these studies have contributed to understanding the nature and causes of falsification, they share some common limitations. The findings are not very reliable and may not generalize to other studies. These studies usually examined a small group of interviewers who were confirmed as, or suspected of, falsifying data in their assignments. The studies also often relied on the reinterview results based on data from a fairly short period of time. Falsification was studied only at the interviewer level, and statistical modeling was rarely undertaken. Some of the proposed methods are effective, but may be difficult to automate and execute in the field.

This article directly incorporates modeling into the sample design for the reinterview, with the goal of creating a reinterview sample design that has a greater chance of identifying falsified cases. The reinterview program of the Current Population Survey (CPS) is used to examine and evaluate the utility of statistical models for the reinterview sample design. Section 2 describes the current method of selecting reinterview samples from the CPS and summarizes its drawbacks. Section 3 discusses modeling options and the available data sources for the modeling. Three alternative sampling strategies are considered in Section 4. In Section 5, they are evaluated and compared to the current CPS reinterview sample design. Two-fold cross validation and a simulation study are used for the evaluation. We conclude by discussing the advantages and disadvantages of the alternative sampling designs, and some of the operational issues that might arise if an alternative is implemented.

2. Current Population Survey Reinterview Program

The CPS is a multistage, stratified household survey that samples about 60,000 households each month in the United States. The households are interviewed monthly to obtain labor force information, as well as other household and person characteristics. The CPS uses a rotating panel design in which sample households are interviewed for four consecutive months, dropped out of the sample for the following eight months, and return to the sample for the same four calendar months a year later, after which they are replaced by a new panel. The first and fifth interviews are generally done face-to-face, while the others are done by telephone when possible. For more details on the CPS design, see U.S. Census Bureau (2006).

The CPS has an ongoing program of quality control that includes reinterviews with a subsample of the sampled households every month. The reinterview asks respondents a few questions to monitor the work of the interviewers, without repeating the full interview. Approximately two percent of CPS interviews are sampled for reinterview each month. The reinterview cases that we consider in this study are selected using a twostage stratified sample design over a 15-month cycle of interviews. First, interviewers are sampled so that an interviewer is in reinterview between one and four times within a 15-month cycle. Based on the findings in Waite (1993; 1997), inexperienced interviewers – those with less than five years of experience on CPS – are sampled more frequently than experienced interviewers. Next, cases are selected from the interviewers' workload depending upon their interviewing experience on CPS. Five reinterview cases are selected for each inexperienced interviewer, and eight reinterview cases are selected for each experienced interviewer. If an interviewer's workload is less than the desired number, then all of

their eligible cases are sampled for reinterview. The CPS quality control program also reinterviews cases that are selected purposively by the regional offices, but these cases are excluded from this study.

The outcome of the reinterview indicates if the disposition of the case was misclassified, or the interview itself was falsified. The reinterviewer makes a judgment as to whether any discrepancies are caused by interviewer falsification. If so, the suspected falsification is reported to the program supervisor immediately. The program supervisor examines each "suspected falsification" case, reviews the original interview notes and the reinterview notes, and determines whether falsification should be suspected. If the supervisor finds the discrepancies are caused by an instrument error, a reinterviewer error, a respondent error, or a good faith error by the interviewer, the case is cleared. If it cannot be determined conclusively that no falsification occurred, then the supervisor must carry out a thorough follow-up investigation of the interviewer's current assignment. Although most cases that are suspected of falsification are cleared or confirmed, some remain classified as "still being suspected," because no definite decision can be made even after the full investigation.

The current CPS reinterview program does not detect many cases of falsification. From January 2004 to June 2006, more than 51,000 cases were sampled for reinterview, and 45,000 of these were from the random sample. Only 43 (0.09%) of the random sample reinterview cases were confirmed to be falsified. The primary reason for detecting so few cases is that falsification is a rare phenomenon. Simply increasing the sample size for the reinterview program to detect more falsified cases would greatly increase cost and burden on respondents. The approach taken here is to investigate alternative reinterview sample designs that might be more effective at identifying falsified cases.

3. Modeling to Find Falsification

To develop methods for improving the sampling of falsified cases in the reinterview program, we began by exploring models to predict falsification using information from the original interviews as predictor variables and reinterview dispositions as the outcomes. The predictor variables used for the modeling come only from the original CPS interview and data on the interviewers because the reinterview is time-sensitive. The respondents should be reinterviewed within a week or two of the original interview. These two sources are the only data that are available within this time frame.

Once the model is fitted from historical data, the predicted probabilities from the fitted model are computed using the current CPS data. These predicted probabilities can then be used to sample current CPS cases for the reinterview. A constraint is that the cost of the new design should be equal to the cost of the current program. We approximate equal costs by fixing the number of cases to be reinterviewed to be the same as in the current program.

3.1. Data Sources

The available data sources for modeling falsification are the original CPS interview, the reinterview, the interviewer database, and data from the supervisory review and thorough investigation. The reinterview dataset, the supervisory review outcomes, and the investigation results jointly provide information on whether falsification was ever

suspected for reinterview cases, and whether suspected falsification was confirmed, was cleared, or is still pending. The original CPS interview is a very large dataset that contains all questionnaire variables and some paradata. The interviewer data set has only a few variables related to the interviewers' experience on the CPS.

Past literature on the CPS reinterview indicated that the candidate predictor variables for modeling should be chosen from the following types of data: (1) key questionnaire variables such as ones that are related to respondents' labor force status; (2) geographic variables such as region and urbanicity; (3) interviewer characteristics such as experience and supervisory title; and (4) paradata variables such as interview mode (telephone/personal visit), length, outcome (complete/noninterview), and timing.

A pool of 30 different CPS monthly datasets was extracted for the period between January 2004 and June 2006, covering two 15-month reinterview cycles. Each CPS case selected randomly for reinterview was classified into one of five possible outcomes: (1) not suspected of falsification in reinterview (N1); (2) suspected in reinterview, but later cleared (N2); (3) suspected in reinterview, and later confirmed (F); (4) suspected in reinterview, and still suspected (neither cleared nor confirmed) (SS); and (5) suspected in reinterview, but classified as "inconclusive" due to missing information (I).

The predictor variables used to fit a model and explain the reinterview outcomes are: Interview outcome; Interview mode; Duration of interview; Interview date; Month-in-sample; and Interviewer's experience. The specific predictor variables are defined in Section 3.2. Other candidate variables, such as labor force status, region, urbanicity, and whether interviewer is regular or supervisory, were considered but eventually were not used for several reasons. Some of these variables were applicable only to a relatively small subset of cases, some had very skewed distributions, and some were found to have little power to explain falsification in our initial explorations.

Table 1 shows the distribution of reinterview outcomes in the dataset. More than 99 percent of the reinterview cases were never suspected of being falsified (this does not imply that none of these cases were actually falsified). Among those that were ever suspected, about half were later cleared. In total, the percentage of cases with confirmed falsification and still suspected after final investigation is less than a quarter of one percent.

About 20 percent of the suspected cases failed to reach a conclusion. To utilize these cases in the modeling, missing outcomes were imputed for about half of them using reinterview outcomes of the cases done by the same interviewers. If an interviewer had

Table 1. Outcomes CPS reinterview cases from January 2004 to June 2006, by falsification status

Outcomes	Count		Percentage	
Not suspected (N1)	44,458		99.29	
Suspected	319		0.71	
Confirmed (F)		41		0.09
Cleared (N2)		139		0.31
Still suspected (SS)		65		0.15
Inconclusive (I)		74		0.17
Total	44,777		100.00	

ever falsified or was still suspected of falsifying CPS interviews or ones within the framework of another U.S. Census Bureau demographic survey in roughly the same time period, we imputed the outcomes of the inconclusive cases in respect of this interviewer as "confirmed" or "still suspected." If an interviewer had not been found to falsify any interviews during that time period, then the inconclusive cases were imputed as cleared. The rationale for this procedure is that if some cases for an interviewer are suspected of falsification, then cases from other surveys for that interviewer might also be inspected. If any work is found to be falsified, then other investigations might not be completed.

As a result of this process, 23 cases were imputed as cleared, nine as still suspected, eleven as confirmed; 31 cases were not imputed and later dropped from the model fitting because no data on investigations in the time period were available.

3.2. Logistic Regression Modeling

Logistic regression is a natural choice for analyzing this type of categorical outcome data (Agresti 2002). The simplest specification of the outcome is a dichotomy: falsified (1) and not falsified (0) cases. This dichotomy does not fully describe the fact that the outcomes are obtained at different stages of investigation (e.g., initial or thorough review). We explored using ordinal or nested logistic regression to deal with this, but decided that the benefits of these models are greatly reduced because the outcome is so rare that it limits the ability to incorporate multiple predictors at different levels. We also considered using random effects in the model to account for interviewer effects. Interviewers might be likely to falsify more than one case if they decide to falsify. This type of modeling was not undertaken because the number of interviews is very large and the number of cases sampled per interviewer is small; this combination causes problems in estimating the model parameters in random effects models. As a result, we used the simple binary logistic regression.

We label a case as "falsified" if the reinterview outcome is F or SS, and set the binary dependent variable used in the logistic regression model to be $Y_i = 1$. All other cases are called "not falsified" and $Y_i = 0$. By defining the "still being suspected" cases as "falsified" for the modeling we capture more problematic cases of interest, even though they were not eventually confirmed to be falsified.

The predictor vector used in the logistic regression models consists of binary variables that were set equal to zero unless the following conditions held:

1. TELSHORT = 1 if it was a completed telephone interview and it either took less than 10 minutes when month-in-sample was 1 or 5, or took less than 4.5 minutes when month-in-sample was not equal to 1 or 5.
2. TELLONG = 1 if it was a completed telephone interview and TELSHORT = 0.
3. PVSHORT = 1 if it was a completed personal visit interview and it took less than 10 minutes when month-in-sample was 1 or 5, or it took less than 4.5 minutes when month-in-sample was not equal to 1 or 5.
4. PVLONG = 1 if it was a completed personal visit interview and PVSHORT = 0.

As a result, any noninterview (e.g., ineligible cases and refusals) was equal to zero for the set of binaries {TELSHORT, TELLONG, PVSHORT, PVLONG}. It was

thus the reference level for this set of variables and excluded from the model to avoid singularity.

5. EFLG = 1: Interviewed by experienced interviewer.
6. HRMISi = 1, i = 1,2,. . .,7: The ith month a household is in the sample. Month-in-sample equal to 8 is the reference level for the HRMISi variables and not included in the model.
7. IN7DAYS = 1: Interview completed within the interview week that starts from Sunday and includes the 19th day of the month (except in December when it is one week earlier).

Table 2 shows the estimated parameters, standard errors, and p-values from the logistic regression. The model fit statistics are given below the table. The variables TELLONG, PVLONG, EFLG, HRMIS1, HRMIS5, and IN7DAYS are statistically significant at the 95 percent level, and PVSHORT and HRMIS4 are statistically significant at the 90 percent level. While the parameter estimates may be affected by the extremely low rate of falsification (King and Zeng 2001), neither estimating the parameters nor estimating the magnitude of the probability of falsification is the goal of this research. The objective is to produce predicted probabilities of falsification that are useful for designing the sample.

Table 2. Estimated logistic regression parameter estimates for falsification for the full CPS data set

Variables	Estimate	S.E.	p-value
Intercept	−3.891	0.343	<0.0001
TELSHORT	−0.029	0.241	0.904
TELLONG	−1.736	0.286	<0.0001
PVSHORT	0.482	0.281	0.087
PVLONG	−0.887	0.299	0.003
EFLG	−0.523	0.191	0.006
HRMIS1	−1.592	0.455	0.001
HRMIS2	−0.522	0.342	0.127
HRMIS3	−0.516	0.341	0.131
HRMIS4	−0.511	0.292	0.080
HRMIS5	−0.761	0.331	0.022
HRMIS6	−0.258	0.310	0.404
HRMIS7	−0.234	0.310	0.451
IN7DAYS	−0.857	0.273	0.002

Model fit statistics		
Criterion	Intercept only	Intercept and covariates
AIC	1,733.509	1,665.946
SC	1,742.217	1,787.868
−2 Log L	1,731.509	1,637.946

Association of predicted probabilities and observed responses	
Percent concordant	62.9
Percent discordant	18.3
Percent tied	18.8
c	0.723

Note: Estimates are based on 44,746 records among which 126 cases were falsified, with 73 falsified cases in Cycle 1 and 53 in Cycle 2.

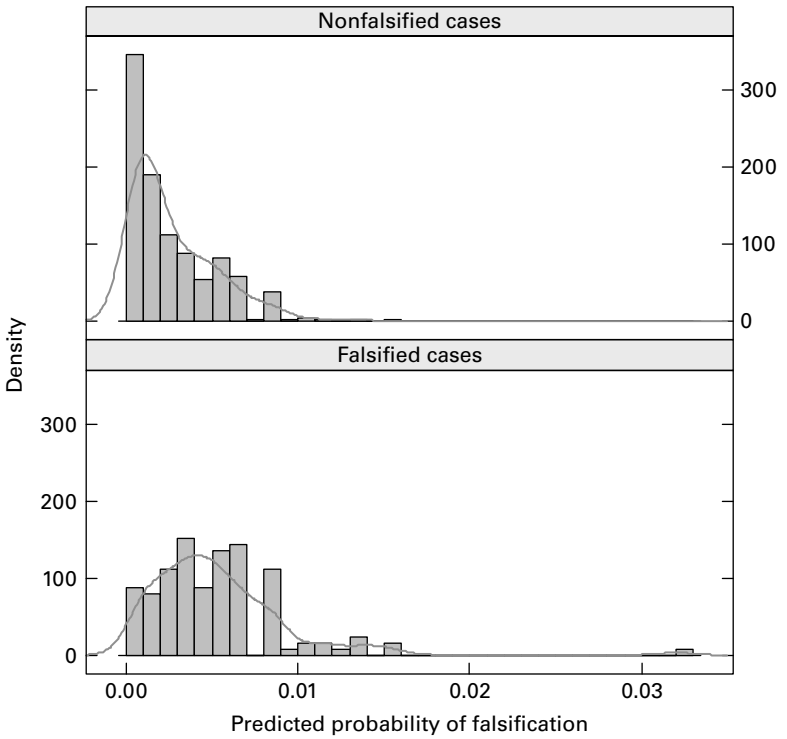

Fig. 1. Histograms and density functions of predicted probabilities of falsification for falsified and nonfalsified cases in the CPS

Figure 1 shows histograms of the predicted probabilities for the falsified and nonfalsified cases. The predicted probabilities were calculated using the formula $\hat{\boldsymbol{p}}_i = \exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)/\left(1+\exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)\right)$ and $\hat{\boldsymbol{\beta}}$ in Table 2. The histogram of the nonfalsified cases is right-skewed, while the falsified case histogram is more uniformly spread. Clearly, the biggest difference is that the nonfalsified cases cluster around a very small predicted probability while the falsified cases do not have this clustering and are more likely to be associated with large predicted probabilities.

4. Alternative Reinterview Sample Designs

The U.S. Census Bureau currently draws CPS reinterview samples using only information on interviewers' experience level. Below, we explore three alternative sample designs that also include data collected in the original interview for reinterview sampling. The alternative designs differ from each other in the way the estimates from the logistic regressions are used in the sampling. Another important difference between the alternatives and the current design is the sampling unit. In the current design, interviewers are sampled and then interviews are subsampled in a second stage. In the alternatives we explore, interviews are directly sampled using characteristics from the cases, bypassing the sampling of the interviewers. Implicit with this approach is the goal of identifying more falsified interviews, rather than interviewers who falsify assignments.

4.1. PPS Design

The first alternative sample design we explore is a Probability Proportional to Size (PPS) design that begins by assigning each original interview a measure of size that is the predicted probability of falsification from a fitted regression model. Interviews are then selected for reinterview with probability proportional to this measure of size. Cases that are more likely to be falsified under the model have a greater chance of selection than those that have lower predicted probabilities of falsification.

The specific sampling procedure uses the estimated parameters from a logistic regression model on historical reinterview data in cycle $t - 1$ to compute the predicted probability of falsification for the original interview case i conducted in cycle t as $\hat{\boldsymbol{p}}_i = \exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)/\left(1 + \exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)\right)$, where $\hat{\boldsymbol{\beta}}$ is the vector of estimated parameters from the model. The predicted probability of falsification is attached to every case that is eligible for reinterview sampling at time t and is used as its measure of size.

4.2. Truncated Design

If the goal is to identify as many falsified cases as possible, an alternative design that does not assign a positive probability of selection to original interview cases that have a low probability of falsification can be considered. We refer to this approach as the truncated design.

The truncated design sorts all the cases that are eligible for reinterview sampling at time t in descending order of the predicted probabilities of falsification. A simple random sample from the cases with the largest probability of selection (e.g., those in the top quintile) is then selected for reinterview. Those cases that are not in the top quintile are not eligible for reinterview sampling. We choose to sample from the top quintile rather than selecting the cases with the largest predicted probabilities to reduce the reliance on the estimated predicted probabilities.

The truncated design should result in sampling more falsified cases for reinterview than the PPS design because it assigns zero probability for those cases with a low predicted probability of falsification. In one sense, the scheme is also less model-dependent than the PPS design since all the model is required to do is to order the cases by their probabilities of falsification. The specific predicted probability is not used in sampling, other than through ranking.

4.3. Stratified Design

Both the PPS design and the truncated design have limitations for a continuous reinterview program that is subject to a variety of changes over time. For example, surveys are subject to external changes such as changes in technology or methods that may affect the falsification of cases. To deal with these types of changes and periodically to verify that the predicted probabilities are reasonably efficient, new modeling should be done to validate the model.

The PPS design samples from all the original interview cases, and thus supports new modeling with standard design-based estimation methods that use the PPS selection weights for reinterview samples. However, the PPS weights used in these analyses may

have a great deal of variation, resulting in unstable design-based estimates with high variances. For example, the ratio of the largest weight to the smallest weight, using the Cycle 1 data, would be 400 to 1.

The truncated design does not support design-based inference because no samples are taken from the pool of the interviews that are classified by the model as having a low probability of being falsified. The truncated approach is, in this sense, less appropriate for an ongoing reinterview program.

A simple compromise is to sample more heavily where most falsified interviews are expected, but sample from the other group and control the selection probabilities to avoid the problems associated with the PPS design. We call this the *stratified* sample design. The predicted probabilities of falsification are used to partition the universe of original interviews into a stratum of those interviews that are most likely to be falsified (the same quintile used in the truncated design) and a stratum of those interviews that are less likely to be falsified. A simple random sample is then selected in each of the two strata.

After considering various alternative allocations, the reinterview sample was allocated in such a way that 75 percent of the sampled cases were from the high probability of falsification stratum and 25 percent were from the low probability of falsification stratum. This allocation results in weights that vary by a factor of twelve, rather than the 400 from the PPS design. The design effect due to differential sampling rates for a mean estimated from such a sample is approximately 2.6. Increasing the allocation to the high probability stratum further would make it possible to select slightly more falsified cases, but results in larger design effects. The stratified design allows reassessment of the model and its parameters that should be reasonably precise.

5. Evaluation

The three alternative sample designs are evaluated below using two different approaches. The purpose of this evaluation is to assess the advantages and disadvantages of the sample designs, not to evaluate the statistical model for predicting probabilities of falsification. To do this, we first use a two-fold cross validation approach to compare the alternatives to each other and the current sampling procedure. Second, we present a simulation study that enables us to consider some properties that could not be assessed using the cross validation approach. The main measure of efficiency for the evaluations is the expected number of falsified cases sampled when the reinterview sample size is held constant. Other criteria such as the ability to reevaluate the model periodically are considered later.

5.1. Twofold Cross Validation

External validation is generally preferred in prediction research, because prediction models tend to perform better on data on which the model was constructed, rather than on new data (Bleeker et al. 2003). The two-fold cross validation is a pseudo-external validation approach that allows us to take advantage of all 30 months of accumulated data from the CPS reinterviews. It fits a regression model from one cycle of reinterviews, uses those estimated model parameters to predict the probabilities of falsification for all original CPS interviews in the other cycle, calculates the selection probabilities for the designs, and then computes the expected number of falsified interviews to be drawn for

reinterview using these designs. The process is then repeated, switching the roles of the data in two cycles.

Let U_c denote the original CPS interviews and S_c denote the reinterviews for Cycle c ($c = 1$ or 2); $S_c \subset U_c$. In U_c the predictor variables $\mathbf{X}_c$ are assumed to be available for all of the original CPS interview cases. The sizes of U_c and S_c are denoted as N_c and n_c. Table 3 shows the sample sizes and the number of experienced and inexperienced interviewers in S_c and U_c that are used in the current design. Since not all interviewers worked all 15 months of the cycle, the number of interviewer-months is given in the table. Interviewer-months are defined as distinct combinations of interviewers and months. If an interviewer conducts CPS interviews in k months within a reinterview cycle, it is counted as k interviewer-months, rather than one.

The detailed steps in the cross validation are given below.

1. Fit the logistic regression model using data in $\boldsymbol{S}_1$ and save the estimated parameters $\hat{\boldsymbol{\beta}}_1$ (for Cycle 1).
2. Apply the estimated parameters $\hat{\boldsymbol{\beta}}_1$ to the data in U_2 and calculate the predicted probability of falsification, $\hat{\boldsymbol{p}}_i = \exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1\right)/\left(1 + \exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1\right)\right)$, $i \in U_2$.
3. Calculate the selection probabilities for each case in U_2 assuming a sample of size $n_2 = 21,821$. Approximate the current approach by a stratified simple random sampling of interviewer-months, with strata depending upon interviewers' experience.
4. Sum up the selection probabilities for the cases that were reinterviewed and were falsified in U_2 (any other falsified cases in U_2 are not known since they were not reinterviewed). Compute the expected yield of known falsified cases to be included in the sample, if a specific sampling design were executed.
5. Repeat Step 1 through 4 by replacing S_1 and U_2 with S_1 and U_1.

The estimated parameters $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ based on Cycle 1 and Cycle 2 data are slightly different, as shown in Table 4. PVSHORT and EFLG are statistically significant in the Cycle 2 model but not in Cycle 1, whereas the significant parameter IN7DAYS in the Cycle 1 model is not significant in Cycle 2. This suggests some instability in the fitted model that may be due to predicting such a rare event or collinearity in the predictors since there were no major changes in CPS operations during the 30 months.

Table 5 gives the sums of the selection probabilities for reinterview cases that were identified as being falsified. If the current sampling design were used to draw reinterview

Table 3. Number of original CPS and CPS reinterview cases, by interviewer-months, interviewer experience, and cycle

	Original interview			Reinterview		
		Interviewer-months			Interviewer-months	
Cycle	Interviews	Experienced	Inexperienced	Interviews	Experienced	Inexperienced
1	897,323	14,301	15,971	22,925	1,599	3,175
2	794,754	14,372	14,301	21,821	1,611	2,935
Total	1,692,077	28,673	30,272	44,746	3,210	6,110

Table 4. Parameter estimates for the two cycles

Variables	Cycle 1 estimate	Cycle 1 *p*-value		Cycle 2 estimate	Cycle 2 *p*-value	
Intercept	−3.798	<0.0001		−4.036	<0.0001	
TELSHORT	−0.104	0.737		0.093	0.807	
TELLONG	−1.801	<0.0001		−1.633	0.000	
PVSHORT	0.102	0.804		0.888	0.026	
PVLONG	−0.826	0.030		−0.951	0.050	
EFLG	−0.294	0.229		−0.841	0.007	
HRMIS1	−1.371	0.017		−1.920	0.012	
HRMIS2	−0.588	0.222		−0.476	0.327	
HRMIS3	−0.184	0.661		−1.143	0.072	
HRMIS4	−0.410	0.298		−0.630	0.152	
HRMIS5	−0.825	0.076		−0.712	0.132	
HRMIS6	−0.262	0.532		−0.260	0.572	
HRMIS7	−0.032	0.936		−0.591	0.255	
IN7DAYS	−0.929	0.008		−0.748	0.091	
	$n = 22{,}925$	falsified cases = 73		$n = 21{,}821$	falsified cases = 53	
Model fit statistics	Criterion	Intercept only	Intercept and covariates	Criterion	Intercept only	Intercept and covariates
	AIC	987.198	964.155	AIC	746.027	719.931
	SC	995.238	1076.715	SC	754.017	831.800
	−2 Log L	985.198	936.155	−2 Log L	744.027	691.931
Association of predicted probabilities and observed responses	Percent concordant	62.2		Percent concordant	62.3	
	Percent discordant	19.7		Percent discordant	15.5	
	Percent tied	18.1		Percent tied	22.2	
	c	0.712		c	0.734	

Table 5. Sum of selection probabilities for the cases identified as falsified in Cycle 1 and 2, by designs

	Cycle 1 (73 known-to-be-falsified cases)	Ratio to current design	Cycle 2 (53 known-to-be-falsified cases)	Ratio to current design
Current design	1.90	1.00	1.76	1.00
PPS design	3.31	1.74	2.57	1.46
Truncated design	3.96	2.08	3.02	1.72
Stratified design	3.31	1.74	2.52	1.43

samples from the Cycle 1 original interviews, then an average of 1.90 of the 73 known-to-be-falsified cases would be included in the reinterview sample. Since we are evaluating the sample designs where the selection probabilities are defined by the predicted probabilities of falsification from the other cycle and those are considered fixed, characteristics such as the mean can be computed without sampling error. The simulation evaluation in the next section avoids the problem of considering the probabilities fixed.

If the PPS design were used, an average of 3.31 known-to-be-falsified cases would be selected. The second column of the table shows that the ratio of the yield from the PPS design to the current design is 1.74 (3.31/1.90). The expected yield is 3.96 for the truncated design, twice the yield of the current design. In Cycle 2, the alternative designs increase the expected yields of falsified cases by 46 and 72 percent. The expected yields of the stratified design are roughly the same as those of the PPS design. This evaluation shows that the alternative designs are much more effective at sampling falsified cases than the current design.

5.2. Simulation Study

A limitation of the cross validation approach is that cases that were not included in the CPS reinterview sample cannot be used since their falsification status is not known. Thus, the yield of falsified cases can only be determined for the cases that were sampled for the reinterview. A second issue is that the modeling treats the samples drawn from the two reinterview cycles as fixed. In fact, these samples are randomly selected, and the sampling variability associated with estimating the model parameters is not accounted for in the cross validation approach. To address these concerns, falsification status was defined for all original interview cases in the simulation. In addition, the reinterview samples used for modeling were drawn independently at each iteration of the simulation.

Table 6. Distribution of the falsification indicator in simulation, by cycle

Y	Cycle 1	Percent	Cycle 2	Percent	Total	Percent
1: falsified	3,178	0.35	2,776	0.35	5,954	0.35
0: not falsified	894,145	99.65	791,978	99.65	1,686,123	99.65
Total	897,323	100.00	794,754	100.00	1,692,077	100.00

In the simulation, the predicted probabilities for each case were generated from a logistic regression, where the predictor vector included an intercept term along with TELLONG, PVLONG, EFLG, HRMIS1, HRMIS5, and IN7DAYS. These predictors were statistically significant at 95 percent level in the model based on combined Cycle 1 and Cycle 2 data. Only the significant predictors were included to allow simple manipulations to test the robustness of the sample designs when the model did not hold. The parameters $\boldsymbol{\beta}$ of the function are taken from Table 2 as fixed constants for the simulation. Next, the falsification indicator, Y_i, $i \in U_1, U_2$, was generated from a Bernoulli distribution using the predicted probability of falsification from the model. Table 6 shows the resulting distribution of Y_i's by cycle. While the distribution of the predictors varies across the cycles somewhat, 0.35 percent of all the cases are falsified in both Cycle 1 and Cycle 2. This is contrary to the actual outcomes that were not as smooth (see Table 4).

Once the falsification status for each CPS original interview case was fixed, the simulation proceeded by drawing a reinterview sample S_1 from Cycle 1 data U_1 using the current sampling scheme, approximated by a stratified simple random sample of interviewer-months as described above. The sampling rate was ten percent for experienced-interviewer-months, and 20 percent for inexperienced-interviewer-months. This step replicated the historical selection of the reinterview sample, but reflects the randomness in drawing samples.

Next, we developed three logistic regression models using different sets of predictors (see Table 7). The first model is called the correct model, and it contains the exact set of predictors used to generate the outcomes. The second is called the overspecified model since it contains all the predictors in the correct model plus additional predictors. The third is called the misspecified model as it does not have one of the important predictors, TELSHORT. The misspecified model also contains more predictors than are in the correct model. All three models were fitted based on S_1.

The estimated model parameters were applied to the original interview data from April 2005, a subset of U_2, to calculate the predicted probabilities of falsification. There are 59,891 interview cases in the April 2005 data set, including 221 confirmed falsified cases. Under the current design a sample of about 1,800 cases would be selected for reinterview, so we used this sample size for all of the alternative designs. For the alternative designs, the selection probabilities were calculated based on estimated probabilities calculated from each of the three models from Table 7 and using the known probabilities of falsification.

As with the cross validation evaluation, the selection probabilities were summed up for the falsified cases to give the expected yield. Table 8 shows the simulation results from

Table 7. Three logistic models with different predictors

Model	Predictors
Correct model	TELLONG, PVLONG, EFLG, HRMIS1, HRMIS5, IN7DAYS
Overspecified model	TELSHORT, TELLONG, PVSHORT, PVLONG, EFLG, HRMIS1- HRMIS7, IN7DAYS
Misspecified model	TELSHORT, PVSHORT, PVLONG, EFLG, HRMIS1-HRMIS7, IN7DAYS

Table 8. Simulation results: average number of falsified cases identified in samples selected using different designs

	Expected number of falsified cases	Standard deviation	Ratio to current design
Current design	6.51	NA	1.00
PPS design			
Known p	11.09	NA	1.70
Correct model	11.10	0.50	1.70
Overspecified model	11.11	0.53	1.71
Misspecified model	9.32	0.47	1.43
Truncated design			
Known p	14.89	NA	2.29
Correct model	14.62	0.56	2.25
Overspecified model	14.01	0.94	2.15
Misspecified model	11.74	0.71	1.80
Stratified design			
Known p	12.28	NA	1.89
Correct model	12.13	0.38	1.86
Overspecified model	11.71	0.65	1.80
Misspecified model	10.15	0.49	1.56

1,972 out of the 2,000 iterations; 28 of the iterations were excluded due to problems evaluating the model (e.g., no falsified cases were observed for a group defined by a model predictor).

With the current CPS reinterview design the expected number of falsified cases in the reinterview sample for the simulation is 6.51. This number of falsified cases is used as the basis for comparisons to the alternatives. Table 8 shows the results of the simulations. The table shows the expected number of falsified cases sampled under the design, the standard deviation of the expected yield due to the variability in estimating the model parameters and calculating the predicted probabilities, and the ratio of the expected yield of the alternative design to the current design.

The PPS design increased the expected yield increased by 40 percent to 70 percent, depending on the model used for computing the predicted probabilities. The truncated design also sampled more falsified cases than the current design, with the expected yield at least doubled when the model was not misspecified. The expected yield of falsified cases from the stratified design is similar to the yield from the PPS design. The value of including the correct predictors in the model is evident in the expected yields in the table. The overspecified model is almost as effective as the correct model. The table also shows that knowing the true probability of falsification does not guarantee an increase in the yield of falsified cases when compared to predicting the probability, especially when the model includes the appropriate predictors.

In the simulation we also explored using different thresholds for defining the truncated design (15%, 20%, and 25%). We found the yield of falsified cases increased using a threshold of 15% rather than 20%, but the increase was small.

In the simulation study, the correlation at the interviewer level was assumed to be zero in the true model. As a result, the simulation cannot measure the effect of omitting the

correlation in the model fitting on the quality of predicted probability of falsification or on the identification of falsification. Shreiner et al. (1988) found that the correlation at the interviewer level is low, but they argued that this requires future research since they had a small sample. We did not study the interviewer effect because the number of reinterviewed cases per interviewer was small.

6. Discussion

The two-fold cross validation evaluation and the simulation study showed that using any of the alternative designs in sampling for reinterviews should identify more falsified cases in the CPS. These increases could be significant, and might have important implications for improving data collection quality in the CPS. The evaluations found that the truncated design is expected to sample about 15 to 20 percent more falsified interviews than the PPS design or the stratified design. Similar gains were likely even if the statistical model was not correctly specified.

When the sampling decision is placed in the broad context of the CPS and its data collection program, the truncated design has some disadvantages. It samples only from the original interview cases that are above a threshold in terms of their predicted probability of falsification, and this limits the ability to reevaluate the model periodically. Since the CPS and virtually all surveys are subject to changes in technology or methods, this is a serious concern.

The stratified design was introduced to compromise between the goals of including more falsified cases in the reinterview sample, and of ensuring a program that could be reevaluated periodically and could be used to estimate the level of falsification. The stratified design yields approximately the same expected number of falsified cases as the PPS design, but it does not suffer from the highly variable weights of the PPS design. We believe these features make the stratified design a better alternative for a continuous survey like the CPS.

While the stratified design has important benefits, it and the other alternative designs also would require major changes in operations. Perhaps the biggest operational challenge is revising the reinterview sampling scheme to incorporate data from the original interview in a timely manner. Any data needed for reinterview sampling must be captured from the interview and transmitted for reinterview sampling, and then the cases sampled for reinterview have to be sent back to the field for data collection. All of this must be done in time so the reinterview can take place shortly after the original interview. This is a challenge that will require development and testing. One simple approach is to use Poisson sampling, but that requires accepting some variability in the reinterview sample size. The U.S. Census Bureau is considering this and alternative approaches to deal with these issues as it revises its reinterview program.

Another feature of the alternative sampling scheme that needs to be examined carefully is the reinterview caseload distribution at different levels of geography and by interviewer. Some of these factors may be easily handled by using appropriate sampling methods, but new approaches need to test to ensure that unanticipated outcomes are not generated.

7. References

Agresti, A. (2002). Categorical Data Analysis. New York: John Wiley and Sons, Inc.

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

Bennett, A.S. (1948a). Survey on Problems of Interviewer Cheating. International Journal of Opinion and Attitude Research, 2, 89–96.

Bennett, A.S. (1948b). Toward a Solution of the Cheater Problem Among Part-time Research Investigators. Journal of Marketing, 2, 470–474.

Bleeker, S., Moll, H.A., Steyerberg, E.W., Donders, A.R.T., Derksen-Lubsen, G., Grobbee, D.E., and Moons, K.G.M. (2003b). External Validation is Necessary in Prediction Research: A Clinical Example. Journal of Clinical Epidemiology, 56, 826–832.

Boyd, H.W. and Westfall, R. (1955). Interviewers as a Source of Error in Surveys. Journal of Marketing, 19, 311–324.

Bushery, J.M., Reichert, J.W., Albright, K.A., and Rossiter, J.C. (1999). Using Date and Time Stamps to Detect Interviewer Falsification. Proceedings of the American Statistical Association, Survey Research Methods Section, 316–320, Alexandria, VA.

Crespi, I.P. (1945). The Cheater Problem in Polling. Public Opinion Quarterly, 431–445.

Evans, F.B. (1961). On Interviewer Cheating. Public Opinion Quarterly, 25, 126–127.

Hood, C.C. and Bushery, J.M. (1997). Getting More Bang from the Reinterview Buck: Identifying "at risk" Interviewers. Proceedings of the American Statistical Association, Survey Research Methods Section, 820–824, Alexandria, VA.

King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9, 137–163.

Krejsa, E.A., Davis, M.C., and Hill, J.M. (1999). Evaluation of the Quality Assurance Falsification Interview Used in the Census 2000 Dress Rehearsal. Proceedings of the American Statistical Association, Survey Research Methods Section, 635–640, Alexandria, VA.

Murphy, J., Baxter, R.K., Eyerman, J., Cunningham, D., and Barker, P. (2004). A System for Detecting Interviewer Falsification. Proceedings of the American Statistical Association, Survey Research Methods Section, 4968–4975, Alexandria, VA.

Schäfer, C., Schräepler, J., Müller, K., and Wagner, G. (2004). Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methods. Proceedings of the American Statistical Association, Section on Survey Research Methods, 4318–4325, Alexandria, VA.

Schräepler, J. and Wagner, G. (2005). Characteristics and Impact of Faked Interviews in Surveys – An Analysis of Genuine Fakes in the Raw Data of SOEP. Allgemeines Statistisches Archiv, 89, 7–20.

Schreiner, I., Pennie, K., and Newbrough, J. (1988). Interviewer Falsification in Census Bureau Surveys. Proceedings of the American Statistical Association, Survey Research Methods Section, 491–496, Alexandria, VA.

Sheatsley, P.B. (1951). An Analysis of Interviewer Characteristics and Their Relationship to Performance, Part II. International Journal of Opinion and Attitude Research, 5, 79–94.

Swanson, D., Cho, M.J., and Eltinge, J. (2003). Detecting Possibly Fraudulent or Error-prone Survey Data Using Benford's Law. Proceedings of the American Statistical Association, Section on Survey Research Methods, 4172–4177, Alexandria, VA.

U.S. Census Bureau. (1993). Falsification by Field Representatives 1982–1992, Memorandum from Preston Jay Waite to Paula Schneider, May 10.

U.S. Census Bureau. (1997). Falsification Study Results for 1990–1997, Memorandum from Preston Jay Waite to Richard L. Bitzer, May 8.

U.S. Census Bureau. (2006). The Current Population Survey Design and Methodology, Technical Paper 66, Washington, D.C. Accessed at http://www.census.gov/prod/2006pubs/tp-66.pdf on August 17, 2009.

Received August 2009
Revised July 2010