Title

Correcting unit nonresponse via response modeling and raking in the California Tobacco Survey

Authors
Berry, C
Flatt, S
Pierce, JP

Publication Date
2023-12-11

Peer reviewed

Journal of Official Statistics, Vol. 12, No. 4, 1996, pp. 349–363

Correcting Unit Nonresponse via Response Modeling and Raking in the California Tobacco Survey

Charles C. Berry,[1] Shirley W. Flatt,[1] and John P. Pierce[1]

The California Tobacco Surveys are the principal data sources used to gather basic information about smoking behavior and to assess the California Tobacco Control Program – the largest program ever undertaken to reduce the impact of tobacco on society. These surveys use a two-phase design; an initial contact with a household – the 'screener' interview – gathers limited information about persons in the household, including their smoking statuses. Subsequent 'extended' interviews gather detailed information from all adult smokers and a sample of non-smokers. In 1992, unexpected results prompted investigation of unit nonresponse among these selected persons. Both nonresponse and smoking status vary according to education and sex, and nonresponse also depends on the age by smoking status cross-classification. A consequence of this and the sub-sampling of non-smokers is that the proportion of individuals interviewed in the 'extended' phase has a complicated pattern depending on age, education, sex, and smoking status. A generalized raking procedure was used to take these patterns into account and was implemented as a set of survey weights. This procedure seemed to reduce bias and variance in estimates of smoking prevalence and to slightly inflate the variance of estimates for variables which are not highly correlated with smoking status.

Key words: Survey weights; auxiliary information; missing data; double sampling; generalized raking.

1. Introduction

In November 1988, California voters mandated the start of the California Tobacco Control Program: the largest and most comprehensive program ever undertaken to reduce the impact of tobacco on society. The Program instituted an additional 0.25 USD per pack tax on cigarettes, and between 1988 and 1993 spent approximately 600 USD million to promote the achievement of a smoke free society. The California Tobacco Surveys (CTS) are the principal data sources used in a series of reports assessing the impact of this program (Burns and Pierce, 1992; Pierce et al., 1993; Pierce et al., 1994). By the end of 1993, three major surveys of households (in 1990–91, in 1992, and in 1993) had been completed. In addition, interviews with adults, adolescents, and pregnant women living in these households were conducted, and a panel survey (using a sub-sample of adults from the 1990–91 survey) was completed.

[1] Department of Family and Preventive Medicine, University of California, San Diego, La Jolla, CA 92093-0622, U.S.A.
Acknowledgement: The work was supported by Contract Number 89-97872 under the Tobacco Control Section of the California Department of Health Services. The authors thank the Chief Editor and two anonymous reviewers for helpful suggestions and criticisms of an earlier draft.

The CTS uses a two-phase procedure for sampling and interviewing. The first phase is a survey of households during which information is obtained about the individuals who live in the household. The second phase is a survey of selected adults who live in the household. This article focuses on the application of a type of post-stratification in the 1992 CTS in which raking is performed on totals derived from the initial household survey as well as on totals derived from external sources. The principal motivation for the procedure used was to reduce bias in key estimates, but the procedure also impacts on variance, reducing the variance of some estimates and increasing that of others.

Section 1.1. reviews the methods used, and Section 1.2. describes some events which resulted in implementation of a generalized raking procedure. Section 2 reviews double sampling for nonresponse and introduces a generalized raking procedure. Section 3 suggests suitable margins for raking. Section 4 gives some of the details of the implementation of the raking procedure actually used and examines empirical features of the weights created by this method. Finally, Section 5 comments on the usefulness of this approach in the context of the California Tobacco Surveys.

1.1. CTS methods

Data were collected via random-digit dialed telephone interviews in a multi-stage sampling procedure (Waksberg, 1978; Brick and Waksberg, 1991). The state was divided into 18 geographic 'regions' consisting of the ten largest counties and eight groups of smaller, contiguous counties. In each 'region', an initial sample of telephone prefixes (each including 100 possible telephone numbers) was screened to yield a set of 100 numbers likely to include residential telephones. These telephone prefixes constituted the primary sampling units (PSUs) for these surveys; a total of 2,495 PSUs were used with the number allocated to each region being approximately proportional to the population of the region. Telephone contact was attempted for a sample of ten numbers in each PSU. An interview was attempted with each residence contacted.

The initial contact with a household resulted in a "screener" interview with an adult respondent (age 18 years and older) who reported the age, sex, race, education, and smoking status (current smoker, former smoker for fewer than five years, never smoked during the past five years) of each individual in the household. In 1990–1991 and in 1992, subsequent "extended" interviews were attempted with each adult classified as a 'current' or 'former' smoker. For those adults reported as 'never' smokers a sub-sample was drawn by performing a Bernoulli trial for each 'never' smoker with $p = 12/43$ and performing an interview if the trial succeeded. It was anticipated that this would result in approximately equal numbers of 'current', 'former', and 'never' smokers in the 'extended' interviews – reflecting a rough balance between the desire for detailed information on 'current' and 'former' smokers and the need to characterize the entire adult population. The 'extended' interview included a detailed classification of current smoking behavior, a lifetime history of smoking behavior, and other topics. Usually, the "extended" interviews were completed on the same day or within a few days of the "screener" interview.

For some persons, complete data were not available for each of the variables used in the raking procedures. For example, in 1992 there were 29,438 persons in the screener survey. For the purpose of developing post-stratification weights, it was necessary to impute age for 45 persons, race for 126 persons, and education for 668. In the adult survey with 7,263 persons, age had to be imputed in five cases, education in 36 cases, and race in 57 cases. This imputation was carried out via a ''hot deck'' procedure as follows: for an individual missing age, one of a group of individuals of the same sex, race, education, and (if possible) smoking status was sampled and that person's age used as the imputed age. Similarly, if race was missing a person of the same sex, age, education, and (if possible) smoking status was sampled. For a person missing education, a person of the same sex, age, race, and (if possible) smoking status was sampled.

Two sets of individual level weights were created – one for the persons reported in the ''screener'' survey and one for the adults responding to the ''extended'' survey. Both sets used a preliminary set of weights, which reflected the sampling probability of each household. These preliminary weights were calculated from the estimated number of residential telephones in the PSU and the number of telephone lines in the household. These household weights were trimmed to avoid variance inflation due to any PSU or any household having an extremely large weight, which probably induces a small bias in estimators based on the resulting weights (see Brick and Waksberg, 1991). In particular, using the subscript h to index the region of the State, i for the PSU number within a region, and j for the household within a PSU, let n_h be the number of PSUs in region h in which one or more households was contacted, and let m_{hi} be the number of households contacted in PSU i of region h. The average number of households contacted per PSU for region h is $\bar{m}_{h.} = \Sigma_{i=1}^{n_h} m_{hi}/n_h$ and the trimmed PSU weight is given by $C_{hi} = min(3, \bar{m}_{h.}/m_{hi})$, where the function $min(\cdot)$ is the minimum. Let $q_{hij} = 1/2$ if household j in PSU i of region h had more than one telephone number and 1 otherwise. The trimmed household weight is given by $H_{hij} = min\ (C_{hi}q_{hij}, 3\Sigma_{i,j}C_{hi}q_{hij}/(\bar{m}_{h.}n_h))$.

The weights for the screener survey were obtained by first assigning to each person the weight for his/her household and raking these weights to statewide totals using a margin for region of residence and sex and a margin for age, race, and educational attainment using the Deming and Stephan (1940) algorithm. Let $\tilde{U} = \{1, \ldots, k, \ldots, K\}$ be the persons in the State and $U \subset \tilde{U}$ be the adults. Also, let $\tilde{S}_1$ denote the sample of persons in the screener and z_k $(k \in \tilde{S}_1)$ be their weights. Let $S_1 \subset \tilde{S}_1$ be the adults in that sample and $S_2 \subset S_1$ be the subjects in the adult extended survey. The notation $\Sigma_{S_2} a_k$ is a shorthand for $\Sigma_{k \in S_2} a_k$.

If person k belongs to household hij, then $z_k^{(0)} = H_{hij}$ is the initial weight assigned to that person. Let $\tilde{\mathbf{x}}_{k1}$ be a row vector consisting of 35 zeroes and a single one indicating to which of the 36 categories of region (18 categories) by sex (2 categories) person k belongs, and let $\tilde{\mathbf{x}}_{k2}$ be a similar row vector indicating to which of 44 categories of race (four categories) by age (five categories) by education (three categories for persons 18 and older) person k belongs. (Since few persons less than age 18 years have advanced education, $\tilde{\mathbf{x}}_{k2}$ was collapsed over education for persons under 18 years – reducing 60 possible categories to 44.) The statewide margins for region-sex and for race-age-education are given by $\tilde{\mathbf{t}}_{x_1} = \Sigma_{\tilde{U}}\tilde{\mathbf{x}}_{k1}$ and $\tilde{\mathbf{t}}_{x_2} = \Sigma_{\tilde{U}}\tilde{\mathbf{x}}_{k2}$. The margins were developed from the March 1990 Current Population Survey and the 1990 Census of Population and Housing. Using ''/'' to denote elementwise division of array elements and ''x''' to denote the transpose of x, the steps in the raking procedure are given by

$$z_k^{(2i_1+i_2)} = \left(\tilde{t}_{x_{i_2}} / \sum_{\tilde{S}_1} z_k^{(2i_1+i_2-1)} \tilde{\mathbf{x}}_{ki_2}\right) \tilde{\mathbf{x}}'_{ki_2} z_k^{(2i_1+i_2-1)}, \qquad i_2 = 1,2$$

These steps are carried out until $max(|z^{(2i_1+2)} - z^{(2i_1)}|) < 10^{-9}$.

For the extended survey weights, v_k $(k \in S_2)$, the same household weights were increased for selected never smokers by a factor of 43/12, i.e., $v_k^{(0)} = H_{hij}$ is the initial weight assigned if adult k is in household hij and was a 'current' or 'former' smoker, while $v_k^{(0)} = 43/12 \times H_{hij}$ is the weight assigned if the adult was a 'never' smoker. The indicator vectors for region-sex remained the same, i.e., $\mathbf{x}_{k1} = \tilde{\mathbf{x}}_{k1}$, but the race-age-education classification was changed once to remove categories for children and again because the raking algorithm experienced convergence problems. The 1992 extended adult survey involved only 7,263 adults and some rather small counts occurred in the table of race by age by education. To correct this, the two lowest categories of education were combined in forming the indicator vectors for race by age by education, $\mathbf{x}_{k2}$, resulting in 24 categories. An additional set of indicator vectors, $\mathbf{x}_{k3}$, for the original three categories of education was used – adding a third step to each cycle of the raking procedure. The totals used for raking were $\mathbf{t}_{x_1} = \Sigma_U \mathbf{x}_{k1}$, $\mathbf{t}_{x_2} = \Sigma_U \mathbf{x}_{k2}$, and $\mathbf{t}_{x_3} = \Sigma_U \mathbf{x}_{k3}$ and the steps of the raking procedure were:

$$v_k^{(3i_1+i_2)} = \left(\mathbf{t}_{x_{i_2}} / \sum_{S_2} z_k^{(3i_1+i_2-1)} \mathbf{x}_{ki_2}\right) \mathbf{x}'_{ki_2} v_k^{(3i_1+i_2-1)}, \qquad i_2 = 1,2,3$$

Thus, a population total for all persons, $\tilde{t}_y = \Sigma_{\tilde{U}} y_k$, would be estimated by $\Sigma_{\tilde{S}_1} y_k z_k$ while a total for adults, $t_y = \Sigma_U y_k$ would be estimated as $\Sigma_{S_1} y_k z_k$ using data from the screener survey or as $\Sigma_{S_2} y_k v_k$ using data from the adult extended survey.

Variance estimates are based on a system of replicate weights using a jackknife procedure. The particular jackknife used is described briefly by Wolter (1985, p. 183) and is a member of the family of "combined strata grouped jackknife" variance estimators studied by Rust (1986). It provides reasonably efficient estimates of variance at a low computational cost. Ideally, primary sampling units (PSUs) in the sample are divided into groups so that each group includes PSUs from all of the regions (i.e., strata), and the fraction of PSUs from any region for any group equals that fraction for the full set of PSUs. Fifty-one such groups were used – resulting in fifty-one sets of replicate weights. Since the numbers of PSUs in each region were not all multiples of 51, it was necessary to relax the condition that the number of PSUs contributed by a region be equal for every group. This will tend to slightly inflate the estimated variance. Fifty-one jackknife samples were formed – each including the PSUs in 50 of the 51 groups. For each jackknife sample, the imputations of age, sex, race, and education were carried out using only those subjects included in that jackknife sample. The screener base weights, $z_k^{(0)}$, were calculated for subjects in that jackknife sample and raked to population totals for region by sex and race by age by education. The weights for subjects not included in that jackknife sample were set to zero. Similarly, the adult 'extended' base weights, $v_k^{(0)}$, for subjects in that jackknife sample were calculated and raked to population totals for region by sex and race by age by education (in two categories) and education (in three categories).

The variance estimate for an estimated population total, $\hat{y} = \Sigma_{\tilde{S}_1} y_k z_k$, is obtained by calculating

$$\hat{y}^{[j]} = \sum_{\tilde{S}_1} y_k z_k^{[j]}$$

for each jackknife sample, j, where $z_k^{[j]}$ is the replicate weight for subject k in sample j. Variance estimates are then obtained as

$$\hat{V}(\hat{y}) = \sum_{j=1}^{51} (\hat{y}^{[j]} - \hat{y})^2 \times 50/51$$

1.2. Hints of a problem in the 1992 adult survey

The estimate of smoking prevalence for adults obtained from the screener survey is given by $\hat{\bar{y}}^{(ss)} = \Sigma_{S_1} z_k y_k^{(ss)} / \Sigma_{S_1} z_k$, where $y_k^{(ss)}$ is a zero-one indicator of whether person k was a current smoker according to the screener interview. The estimate obtained from the 'extended' survey is given by $\hat{\bar{y}}^{(es)} = \Sigma_{S_2} v_k y_k^{(es)} / \Sigma_{S_2} v_k$, where $y_k^{(es)}$ is a zero-one indicator of whether person k was a current smoker according to the extended interview. In the 1990–91 survey, these estimates were in close agreement. This might have been expected in view of the high degree of concordance between the $y_k^{(ss)}$ and the $y_k^{(es)}$ and the similarity of the methods used to form z_k and v_k. However, in 1992 the agreement was not good with $\hat{\bar{y}}^{(ss)} = 19.7\%$ and $\hat{\bar{y}}^{(es)} = 21.8\%$. Using an alternative estimate based on the adult extended respondents but using the screener smoking classification, $\tilde{\bar{y}}^{(ss)} = \Sigma_{S_2} v_k y_k^{(ss)} / \Sigma_{S_2} v_k$, a prevalence of 21.9% was obtained. This difference of 2.2% seems much too large to be explained by chance variation involved in selecting respondents for the adult extended survey or in ignorable self selection. When only females are considered, the difference is more than 3.75%, which again is much too large to be explained by chance. Still another estimate of smoking prevalence was obtained using all of the individuals who were selected for inclusion in the adult extended survey. A set of weights like v_k was developed for those individuals by the raking method used for the extended interview. An estimate of $\bar{y}^{(ss)}$ of 19.9% was obtained which is in reasonable agreement with the weighted screener prevalence of 19.7%. Thus, the method of weighting produced an estimate similar to that in the screener survey when applied to the intended adult extended survey respondents, but not when applied to those who actually responded. This suggested that differential unit nonresponse for different types of respondents might have biased the estimate of smoking prevalence in the 1992 adult extended survey.

There were suggestions that pattern of nonresponse was different in the 1992 surveys. The fraction of "screener" interviews per household statewide declined to 73.1% in 1992 from 75.1% in 1990–1991, while the fraction of "extended" interviews per selected adult statewide declined to 69.9% in 1992 from 75.3% in 1990–1991. The 1992 CTS was fielded in the months March through July of 1992 – a period encompassing the riots in Los Angeles following the "not-guilty" verdict in the trial of white police officers who were videotaped severely beating Rodney King, an African-American motorist. These riots resulted in more than 50 deaths and billions of dollars of property losses. Survey supervisors reported increased difficulty in obtaining interviews in Los Angeles. It seems plausible that these events influenced response rates, although we had no direct evidence of this.

We felt obliged to study the method of estimation for the 1992 adult extended survey to determine if there was a problem in the method of weighting and estimation – possibly related to patterns of nonresponse – and if such a problem were found to take corrective action. Moreover, we felt that the form of any correction should be a new set of weights; each of these datasets is released for public use after the publication of an initial report, and it was felt that public users should be able to reproduce exactly the results that appear in the report without needing access to (or knowledge of) special estimation procedures. It turned out that there was a problem with the method of weighting and estimation related to differential nonresponse among subgroups of subjects. These differentials are explored in Section 3.

2. Double Sampling for Nonresponse

Since only a fraction of the 'never' smokers identified in the screener survey are selected for the extended interview, the design is a two-phase or double-sampling design. Double-sampling designs date from Neyman's (1938) estimator which stratifies the second phase observations according to totals obtained in the first phase and Cochran's (see Watson 1937) regression estimator which corrects the second phase total according to the mean of an auxiliary variable measured in the first phase sample. Sometimes double sampling arises because of nonresponse (rather than intentional subsampling in a second phase) causing bias in estimation. Nowadays, nonresponse is viewed as a random event – in contrast with the classical view of nonrespondents as members of a separate stratum – and models of the response mechanism are constructed as a means of reducing bias. Särndal and Swensson (1987) proposed regression estimators for use in settings with nonresponse in the second phase. Their simulation results showed that even if the assumed response model is incorrect, using auxiliary variables which are highly correlated with the variable of interest greatly reduces bias.

2.1. *Calibration estimators*

Several treatments of nonresponse and double sampling in a finite population setting have been based on regression estimators (including Folsom 1981; Mukerjee and Chaudhuri 1990; and Särndal and Swensson 1987). The notions of *generalized raking* and *calibration estimation* are useful in treating double sampling, nonresponse adjustment for bias-reduction, and post-stratification in the present framework. Using the notation of the adult 'extended' interview survey, let $\mathbf{x}_k = (x_{k1}, x_{k2}, x_{k3})$ be a row vector which combines all of the auxiliary information for subject k and let $\mathbf{t}_x = (t_{x_1}, t_{x_2}, t_{x_3})$ be the corresponding vector of population totals. A calibration estimator is based on a distance function, $G(v_k/v_k^{(0)})$, which measures the distance from the original weight, $v_k^{(0)}$ (which is often the inverse sampling probability), to the new weight, v_k, and a set of side conditions $\mathbf{t}_x - \Sigma_{S_2} v_k \mathbf{x}_k = 0$. The problem of minimizing $\Sigma_{S_2} v_k^{(0)} G(v_k/v_k^{(0)})$ subject to the constraints leads to a set of *calibration equations*, $\Sigma_{S_2} v_k^{(0)} F(\mathbf{x}_k \lambda)\mathbf{x}_k = \mathbf{t}_x$, where $F(\cdot)$ is the inverse function of $\partial G(u)/\partial u$. These are solved for λ, the vector of unknown Lagrangian multipliers, and then $v_k = v_k^{(0)} F(\mathbf{x}_k \lambda)$. Deville and Särndal (1992) noted that the weights obtained by solving calibration equations have an appealing property: if they are applied to an auxiliary variable, they must give perfect estimates, so when applied to a variable that is strongly correlated with an auxiliary variable, they should perform well. Deville and Särndal (1992) introduced a class of distance measures which includes special cases leading to generalized regression estimators, classical raking, and trimming of weights. As one example, Deville and Särndal (1992) show that the generalized raking procedure with distance function $G(u) = u^2 - 1$ is the generalized regression estimator of Cassel, Särndal, and Wretman (1976). When the inverse function, $F(\mathbf{x}_k\lambda) = 1 + \mathbf{x}_k\lambda$, is substituted into the calibration equation and solved for λ we have

$$\lambda = \left(\sum_{S_2} v_k^{(0)} \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \left(\mathbf{t}_x' - \sum_{S_2} v_k^{(0)} \mathbf{x}_k'\right)$$

and further substitution shows that this leads to the generalized regression estimator

$$\hat{t}_y = \sum_{S_2} v_k y_k = \left(\mathbf{t}_x - \sum_{S_2} v_k^{(0)} \mathbf{x}_k\right) \left(\sum_{S_2} v_k^{(0)} \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \left(\sum_{S_2} v_k^{(0)} \mathbf{x}_k' y_k\right) + \sum_{S_2} v_k^{(0)} y_k$$

The term *generalized raking procedure* was used by Deville et al. (1993) for procedures in which two or more sets of marginal counts are used to define the side conditions.

2.2. *Calibration to two levels of auxiliary information*

Calibration estimators can be set up to use information available in the first phase sample only as well as known population totals. Consider a set of side conditions which incorporates information available in the first phase sample, but for which the population totals are unknown, $\hat{\mathbf{t}}_r - \sum_{S_2} v_k \mathbf{r}_k = 0$ where $\mathbf{r}_k$ is a vector of variables available for each member of the first phase sample and $\hat{\mathbf{t}}_r$ is the corresponding vector of estimated population totals. Using these along with the other side conditions leads to the calibration equations

$$\sum_{S_2} v_k^{(0)} F(\mathbf{x}_k\lambda_x + \mathbf{r}_k\lambda_r)(\mathbf{x}_k, \mathbf{r}_k) = (\mathbf{t}_x, \hat{\mathbf{t}}_r) \qquad (2.2.1.)$$

and to weights

$$v_k = v_k^{(0)} F(\mathbf{x}_k\lambda_x + \mathbf{r}_k\lambda_r) \qquad (2.2.2.)$$

Some care is required in estimating $\mathbf{t}_r$. These calibration equations will have no solution unless $\hat{\mathbf{t}}_r$ and $\mathbf{t}_x$ satisfy a condition which assures that the estimates in $\hat{\mathbf{t}}_r$ are not contradicted by the population totals in $\mathbf{t}_x$. To wit, for any vectors of constants, L_x and L_r, satisfying

$$x_k L_x = \mathbf{r}_k L_r \qquad k \in S_2, \qquad (2.2.3.)$$

$$\text{then } \mathbf{t}_x L_x = \hat{\mathbf{t}}_r L_r \qquad (2.2.4.)$$

must also be satisfied. This follows since (2.2.3.) gives $\Sigma_{S_2} v_k \mathbf{x}_k L_x = \Sigma_{S_2} v_k \mathbf{r}_k L_r$ which with (2.2.1.–2.2.2.) gives (2.2.4.). One can construct $\hat{\mathbf{t}}_r$ to satisfy the condition as follows: Let $\mathbf{L}_x$ be a matrix whose columns are a basis for those L_x which satisfy (2.2.3). Solve the calibration equation

$$\sum_{S_1} w_k^{(0)} F(\mathbf{x}_k \mathbf{L}_x \lambda_0) \mathbf{x}_k = \mathbf{t}_x \mathbf{L}_x \tag{2.2.5.}$$

and obtain weights $w_k = w_k^{(0} F(\mathbf{x}_k \mathbf{L}_x \lambda_0)$, $k \in S_1$. Choose $\hat{\mathbf{t}}_r$ as $\hat{\mathbf{t}}_r = \Sigma_{S_1} w_k \mathbf{r}_k$. The w_k satisfy $\mathbf{t}_x \mathbf{L}_x = \hat{\mathbf{t}}_x \mathbf{L}_x = \Sigma_{S_1} w_k \mathbf{x}_k \mathbf{L}_x$ assuring that $\mathbf{t}_x L_x = \hat{\mathbf{t}}_r L_r$.

Deville and Särndal (1992) remarked that the weights created by a calibration estimator are as close as possible to the original weights among those that satisfy the side conditions. Thus, recalibration of existing weights, z_k, for a good estimator to a minimal set of constraints would be expected to yield another good estimator. Setting $w_k^{(0)} = z_k$ and solving 2.2.5.) to get w_k gives an estimate of $\mathbf{t}_r$ which satisfies (2.2.3.–2.2.4.) and yet is close to the estimate based on z_k. Another possibility is to let $w_k^{(0)}$ be the sampling weights (e.g., $z_k^{(0)}$) and replace L_x in (2.2.5.) with the identity matrix. This yields a set of weights which respects all of the side conditions in (2.2.1.) as well as (2.2.3.–2.2.4.); several existing double-sampling estimators can be constructed in this way.

A number of estimators proposed for two phase sampling can be viewed as solutions to calibration equations similar to (2.2.1.) and (2.2.5.). A simple dual-stratification estimator arises when $\mathbf{x}_k$ is a set of indicators for the categories of a stratifying variable, $\mathbf{t}_x$ contains the known population counts, $\mathbf{r}_k$ is a set of indicators for a finer set of categories (i.e., $\mathbf{r}_k$ maps to $\mathbf{x}_k$) which is available only for S_1, and $F(u) = \exp(u)$. This is a special case of the family of estimators proposed by Vardeman and Meedan (1984). Mukerjee and Chaudhuri (1990) studied a generalized regression estimator, which would arise when $\mathbf{x}_k$ and $\mathbf{r}_k$ are scalar valued variables observed in the first phase sample, $\mathbf{t}_x$ is a known total for the population, and $F(u) = 1 + u$. They note that if $\mathbf{t}_r$ had been known exactly, it would have been used in forming the regression estimator; since it is not, a regression estimate of $\mathbf{t}_r$ is used instead based on $(\mathbf{x}_k, \mathbf{r}_k)$ for S_1.

The strategy which we used is to perform exploratory analyses to develop a model for the differences between the 1992 CTS adult extended survey respondents and the screener respondents and to use the smoking status from the screener survey as an auxiliary variable in the adult extended survey. The development of the model amounts to choosing sets of totals to be used in (2.2.1.) and specifying the functional form of $F(u)$. The calibration equations will use $F(u) = \exp(u)$; these calibration equations can be solved by applying the classical Deming-Stephan raking algorithm to equation (2.2.5.), then estimating $\hat{\mathbf{t}}_r$, and finally applying the Deming-Stephan algorithm to equation (2.2.1.). In setups in which all sampled subjects respond, choices of $F(u)$ within the class studied by Deville and Särndal (1992) are asymptotically equivalent – suggesting that the choice is of little consequence. However, when nonresponse is present, Little and Wu (1991) show that different choices of $F(u)$ imply different response mechanisms; under multinomial sampling each choice leads to maximum likelihood estimates for the parameters of the implied model. The choice of $\exp(u)$ implies that the logarithm of the response probability is a linear function of the variables in the assumed response model. Little and Wu also performed a simulation study comparing the performance of estimates based on different choices of $F(u)$ in various setups. They noted that the choice $\exp(u)$ performs well even when applied to a setup generated by another function. Based on this simulation, they recommended use of $\exp(u)$ in the absence of knowledge of the mechanism governing nonresponse.

3. Exploring Unit Nonresponse

Table 3.1. shows what fraction of the adults selected for an extended interview actually had such an interview according to the sex, region (dichotomized as LA versus all others), race, age, education, and smoking status from the screener interview.

The original raking procedure for the adult 'extended' survey used a sex-region margin and a race-age-education margin to compensate for differential response rates and under-coverage among the different population subgroups defined by either of these margins. The variation in response according to smoking status is not marked; it might seem that smoking status could be ignored. However, there are two reasons why this expectation is not met. First, the full table formed by crossing unit nonresponse to the extended survey with the other variables might contain patterns of nonresponse which are neither evident in Table 3.1 nor corrected using the existing set of margins for raking. Second, the sampling procedure used in the second phase specifies much different sampling fractions for 'never' smokers ($f = 12/43$) and for all other adults ($f = 1.0$). When there is differential nonresponse which does not depend on screener smoking status (say for males versus females), the second phase sampling yields inclusion probabilities which depend on both sex and smoking status. In general, the odds-ratio among the 'never' smokers tends to be biased towards one. As an example, suppose that the probability of nonresponse for males is .34 and for females is .26 regardless of smoking status; the odds-ratio is 1.47 for nonresponse by sex. Since all smokers are included in the extended survey if they respond, the odds-ratio for inclusion versus sex among smokers is 1.47. The 'never' smokers will be included if they are selected and respond. For the males, the probability is $(1 - .34) * 12/43 = .184$, while for females the probability is $(1 - .26) * 12/43 = .207$; the odds-ratio for inclusion versus sex among 'never' smokers of 1.15 differs markedly

Table 3.1. Unit nonresponse by type of respondent

Variable	Category	Nonresponse rate in per cent
Sex	Male	34
	Female	26
Region	Los Angeles	35
	Other	28
Ethnicity	Non-Hispanic white	26
	African-American	34
	Hispanic	38
	Other	41
Age	18–24	31
	25–44	29
	45 and over	31
Education	0–11 years	40
	12 years	32
	More than 12	26
Screener smoking status	Current smoker	31
	Former smoker	28
	Non-smoker	30

Based on 10,387 individuals selected from the screener survey for the adult extended survey.

from that for smokers. As a consequence, models for the probability of inclusion in the second phase survey must account for the net effects of nonresponse and second phase sampling.

Table 3.2. is based on the fitted probabilities of inclusion in the 'extended' survey. A log-linear model was used to analyze a table formed by crossing inclusion status by smoking status by region by sex by race by age by education of 'screener' sample adults. Terms of successively higher order were added until no other terms attained the nominal level of $p < .05$. The patterns of inclusion are fairly complicated; the odds that a female will be included are greater than for a male, and this tendency is strongest among current smokers. The odds of inclusion according to age and according to education also depend on smoking status.

From this analysis, it would seem that margins for 'smoking status-age', 'smoking status-education', and 'smoking status-sex' should be included in the raking scheme. Särndal and Swensson (1987) have pointed out that bias in estimates can be reduced if powerful explanatory variables are used as auxiliary variables and that these variables have the secondary effect of reducing variance. Taking just screener 'smoking status' as an auxiliary variable should reduce the bias in extended survey variables that are highly correlated with it; since the population total is accurately estimated for the auxiliary variable, totals of variables which are highly correlated inherit much of this accuracy. However, it is also of interest to examine subpopulations of adults; the totals for these subpopulations will be better estimated if auxiliary variables are included which sum to the subpopulation totals. Thus, including 'smoking status-sex' totals may reduce the bias of estimators for male and for female totals better than 'smoking status' alone. It is evident from the construction of generalized raking estimators that as more side conditions are added, the derived weights will tend to depart from the original weights. When the variable of interest is not highly correlated with the auxiliary variable(s) involved in the side conditions, this tends to inflate the variance of the estimated total. Instability in the weights under complete post-stratification – and the attendant inflation of variance – is a common motive for raking to margins of the table defining the post-strata (see Deville et al., 1993 and Little 1993). Thus, the addition of side conditions may improve the estimation of some totals and degrade that of others.

Table 3.2. Relative probability of inclusion according to smoking status

	Current	Former	Never
Sex			
Females (vs males)	1.16	1.11	1.10
Age			
25–44 (vs 18–24)	1.01	1.02	1.35
45+ (vs 18–24)	0.91	0.85	1.02
Education			
12 years (vs <12)	1.08	1.07	0.97
>12 years (vs <12)	1.14	1.09	0.99

The probabilities are based on the fitted model which has 'inclusion status-smoking status-sex', 'inclusion status-smoking status-age', 'inclusion status-smoking status-education', their lower order relatives, and other terms (see text).

So which margins should be added to the original set to obtain less biased estimates of smoking prevalence? Including more margins in the raking scheme has the potential for reducing bias and variance in variables which are correlated with the screener smoking status at the price of increasing the variance of variables which are not highly correlated with screener smoking status. Thus, the addition of margins to the raking scheme should seek a balance between the reduction in variance and bias for some variables and the inflation of variance for others. In what follows, we consider a setup in which only a total for screener smoking status is added to the margins used in the original weighting system for the adult extended survey and another setup in which totals are added for 'smoking status by age,' for 'smoking status by education,' and 'smoking status by sex.'

4. Implementation of Raking Procedures

As noted in Section 2.2, totals derived from the first phase sample must not contradict known totals for the population, and this can be assured by calibrating the first phase totals to certain population totals. This was implemented here by calibrating the screener survey weights, z_k, to the sex and the age-education totals used for the adult extended survey via classical raking and using the calibrated weights to obtain totals for screener smoking status, screener smoking status by sex, etc. The base weights for the extended survey respondents, $v_k^{(0)}$, were then raked using the population totals (i.e., region of residence by sex, age by race by education in two categories, and education in three categories) and either the totals for smoking status or those for 'smoking status by age', for 'smoking status by education,' and 'smoking status by sex.' Table 4.1. shows the combinations of marginal configurations used in each set of weights; the 'original' setup is the one described in Section 1.1, setup A uses just smoking status, and setup B uses 'smoking status by age,' 'smoking status by education,' and 'smoking status by sex.'

Variance estimates are formed by creating sets of replicate weights along the lines of Section 1.1, but with the addition of the calibration step just described. That is, the screener survey weights for every jackknife sample $z_k^{[j]}$ are calibrated to sex and the age-education totals used for the adult extended survey and then used to obtain totals for screener smoking status, screener smoking status by sex, etc. The adult extended base weights, $v_k^{(0)}$, for each jackknife sample are raked to population totals used in the adult extended survey and to either the smoking status totals or the totals for smoking status by sex, etc. Variance estimates are then calculated as described in Section 1.1.

Table 4.1. Marginal configurations used in various raking schemes

	Setup		
Configuration	original	A	B
Region-sex	•	•	•
Race-age-education dichotomy	•	•	•
Education (3 categories)	•	•	•
Smoking status		•	
Sex-smoking status			•
Education-smoking status			•
Age-smoking status			•

Table 4.2. *Some means and variances under different raking schemes*

	Mean			Variance ratio		
	original	A	B	original	A	B
Current smoker	0.221	0.200	0.201	1	0.455	0.461
Msmoke–Fsmoke	0.023	0.024	0.056	1	0.884	0.657
Eats out rarely	0.318	0.319	0.319	1	1.084	1.085
TV hours	2.894	2.895	2.898	1	1.088	1.096
Work for other	0.548	0.548	0.548	1	1.077	1.073
Increase tax	0.440	0.440	0.440	1	1.081	1.087
No smoking	0.400	0.400	0.400	1	1.079	1.115
Causes cancer	0.838	0.838	0.838	1	1.044	1.057

Table 4.2 gives estimates of population proportions for several variables from the 1992 adult extended survey. The row "Current smoker" gives the proportion who are cigarette smokers according to the adult extended questionnaire. This variable is highly correlated with the screener smoking status variable. "Msmoke – Fsmoke" is the difference of the proportion of males who are current smokers and that of females. "Eats out rarely" is the proportion who eat restaurant meals once a month or less frequently. "TV hours" is the average number of hours of television viewing among those who watch TV. "Work for other" is the proportion who work, but are not self-employed. "Increase tax" is the proportion of subjects who responded that cigarette taxes should be increased. "No smoking" gives the proportion of adults living in homes in which smoking is totally forbidden inside the house. "Causes cancer" is the proportion who agree that inhaling smoke from another person's cigarette causes cancer.

As Table 4.2 shows, there is little impact on the estimated values, except for "Current smoker" which is about 2% lower for setups A and B compared to the original setup and "Msmoke – Fsmoke" which is more than 3% higher under setup B than under setup A or the original setup. These are substantial differences; the annual decline in smoking prevalence is typically less than one percent and a fluctuation as large as 2% in a single year might be interpreted as a major shift in smoking behavior. Apparently, setups A and B result in substantially different estimates of smoking status than the original weighting method. Table 4 also gives the ratios of the estimated variances for setups A and B to the variance under the original setup. It is evident that the variance of "Current smoker" is markedly reduced under both 'A' and 'B,' that of "Msmoke – Fsmoke" is reduced under 'A' and further reduced under 'B,' and that the other variances are slightly increased. It should be noted that these other variables have only modest correlations with screener smoking status. Evidently, there is an initial increase in variance due to adding the screener smoking status margin, but little further increase due to adding margins which cross the other variables with screener smoking status. A rough indication of the relative variance under different sets of weights is given by Σv_k^2 for characteristics which are unrelated to the auxiliary variables. The increases in Σv_k^2 for setups A and B compared to the original setup are 8% and 8%. The increases in jackknife estimates of variance in setups A and B approximate these except for "Current smoking," "Msmoke – Fsmoke," "Causes cancer," and perhaps "No smoking" for setup B.

5. Remarks

In this study, it was shown that the subjects who were included in a second phase interview differed from those in the first phase in an unexpected manner. The pattern of inclusion was the result of intentional sampling of only a fraction of those identified in the first phase as 'never' smokers and unit nonresponse in the second phase. This pattern involved smoking status, education, age, and sex. Since a principal interest in the California Tobacco Surveys is to assess smoking related behavior, it is important to accurately estimate totals that are linked to smoking status. Thus, detecting and correcting unusual patterns of inclusion is important. It is fortunate that the first phase or 'screener' survey collected data on smoking status, sex, age, and education, which could later be used to explore inclusion patterns and provide a means for adjusting the results.

The situation in this study differs somewhat from that in many others. Often auxiliary information is available only in the first phase sample (e.g., as in Tenenbein 1970). Occasionally, information may be available on all survey nonrespondents (not just the second phase nonrespondents) allowing direct comparison of the sampled and target populations (e.g., Ekholm and Laaksonen 1991). Here population totals for some variables are available from the Census and from large external surveys, but the sampling method does not identify first phase nonrespondents. The smoking status variable is only available for households which responded to the initial screener survey, so adjustments based on smoking status depend on the quality of the first phase sample and estimation procedures. The 'screener smoking status' and the 'extended smoking status' do not always agree; the latter is derived from a more extensive set of questions and the former is often reported by another family member. Substantial misclassification in the first phase sample is not in itself a problem for double sampling estimators; this is most obviously the case for maximum likelihood double sampling estimators. However, if persons who are misclassified in the first phase have different inclusion probabilities in the second phase, this could induce a bias which our approach might not reduce.

Examination of the actual patterns of unit nonresponse does not reveal an obvious cause for differential nonresponse. Was the nonresponse somehow related to the riots of 1992 in Los Angeles? It really is not clear from the data that we have in hand. Certainly the failure to find a significant difference for Los Angeles compared to the other counties in pattern of nonresponse suggests that those problems that existed there were found outside of Los Angeles as well.

In addition to correcting for bias due to patterns of inclusion in the second phase sample, a benefit of incorporating the auxiliary information in screener level data for this study in the survey weights was a marked decrease in the variance of variables which are highly correlated with the margins which include screener smoking status. The cost of introducing a raking scheme which includes additional side conditions is that the variance of a measurement which is not highly correlated with screener smoking status is increased. In the context of the goals of this study, this cost seems minor.

6. References

Brick, J.M. and Waksberg, J. (1991). Avoiding Sequential Sampling with Random Digit Dialing. Survey Methodology 17, 27–41.

Burns, D. and Pierce, J.P. (1992). Tobacco Use in California 1990–1991. Sacramento: California Department of Health Services.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1976). Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. Biometrika 63, 615–620.

Deming, W.E. and Stephan, F.F. (1940). On a Least-Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. Annals of Mathematical Statistics 11, 427–444.

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association 87, 376–382.

Deville, J.C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association 88, 1013–1020.

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. Journal of Official Statistics 7, 325–337.

Folsom, R. Jr. (1981). The Equivalence of Generalized Double Sampling Regression Estimators, Weight Adjustments, and Randomized Hot Deck Imputations. Proceedings of the Section on Survey Research Methods, American Statistical Association, 400–405.

Little, R.J.A. and Wu, M. (1991). Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ. Journal of the American Statistical Association 86, 87–95.

Little, R.J. (1993). Post-stratification: A Modeler's Perspective. Journal of the American Statistical Association 88, 1001–1012.

Mukerjee, R. and Chaudhuri, A. (1990). Asymptotic Optimality of Double Sampling Plans Employing Generalized Regression Estimators. Journal of Statistical Planning and Inference 26, 173–183.

Neyman, J. (1938). Contribution to the Theory of Sampling Human Populations. Journal of the American Statistical Association 33, 101–116.

Pierce, J.P., Farkas, A., Evans, N., Berry, C., Choi, W., Rosbrook, B., Johnson, M., and Bal, D.G. (1993). Tobacco Use in California 1992. A Focus on Preventing Uptake in Adolescents. Sacramento: California Department of Health Services.

Pierce, J.P., Cavin, S., Macky, C., Rosbrook, B., and Berry, C. (1994a). Technical Report on Analytic Methods and Approaches Used in the Tobacco Use in California, 1992 Report. Sacramento, California Department of Health Services.

Pierce, J.P., Evans, N., Farkas, A., Cavin, S.W., Berry, C., Kramer, M., Rosbrook, B., Choi, W., and Kaplan, R.M. (1994b). Tobacco Use in California. An Evaluation of the Tobacco Control Program in California. Sacramento: California Department of Health Services.

Rust, K.F. (1986). Efficient Replicated Variance Estimation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 81–87.

Särndal, C.E. and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. International Statistical Review 55, 279–294.

Tenenbein, A. (1970). A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications. Journal of the American Statistical Association 65, 1350–1361.

Vardeman, S. and Meedan, G. (1984). Admissable Estimators for the Total of a Stratified Population that Employ Prior Information. Annals of Statistics 12, 675–684.

Waksberg, J. (1978). Sampling Methods for Random Digit Dialing. Journal of the American Statistical Association 73, 40–46.
Watson, D.F. (1937). The Estimation of Leaf Areas. Journal of Agricultural Science 27, 474–483.
Wolter, K.M. (1985). Introduction to Variance Estimation. Springer-Verlag, New York.

Received February 1994
Revised February 1996