Article

Cell collapsing in poststratification

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2007
Vol. 33, No. 2, pp. 139-150
Statistics Canada, Catalogue o. 12-001-X

Cell collapsing in poststratification

Jay J. Kim, Jianzhu Li and Richard Valliant [1]

Abstract

Poststratification is a common method of estimation in household surveys. Cells are formed based on characteristics that are known for all sample respondents and for which external control counts are available from a census or another source. The inverses of the poststratification adjustments are usually referred to as coverage ratios. Coverage of some demographic groups may be substantially below 100 percent, and poststratifying serves to correct for biases due to poor coverage. A standard procedure in poststratification is to collapse or combine cells when the sample sizes fall below some minimum or the weight adjustments are above some maximum. Collapsing can either increase or decrease the variance of an estimate but may simultaneously increase its bias. We study the effects on bias and variance of this type of dynamic cell collapsing theoretically and through simulation using a population based on the 2003 National Health Interview Survey. Two alternative estimators are also proposed that restrict the size of weight adjustments when cells are collapsed.

Key Words: Bias; Combining cells; Coverage error; Poststratification; Under-coverage; Weight trimming.

1. Introduction

Poststratification is a common technique used in survey weighting that can serve to (1) reduce variances or (2) adjust for deficient coverage by the sample of some groups in the target population. In household surveys in the U.S. the second purpose is especially important because some demographic groups, like young Black males, are covered less well than others (*e.g.*, see Kostanich and Dippo 2000, chapter 16). Adjusting for undercoverage can lead to differential weights, which may correct for bias but will also increase standard errors. Practitioners often avoid making extreme weight adjustments, in effect trading-off some bias reduction in order to keep variances under control.

One method of controlling the size of weight adjustments is to collapse the initial poststratification cells together if the adjustment in a cell exceeds some limit. Little (1993) and Lazzeroni and Little (1998) cover methods of collapsing categories of ordinal poststratifiers. Other strategies for how to collapse strata or construct estimators have been suggested by Fuller (1966), Kalton and Maligalig (1991), and Tremblay (1986). Kim, Thompson, Woltman, and Vajs (1982) give some practical applications. In this paper, we study the effects on bias and variance of combining cells, assuming that more finely defined cells would be preferable if the sample sizes and sizes of weight adjustments were within some tolerances set by the survey designers.

Two criteria are often used to decide whether a cell should be collapsed with another. The first is the *inverse coverage ratio or initial adjustment factor* (IAF), and is defined as the ratio of the control count to the initially weighted sample count for the cell. A ratio which is significantly different from 1 indicates that coverage is either low or high for the group represented by the cell. When the IAF for a cell falls outside some bounds set in advance, the cell is combined with another. For example, the collapsing threshold for "high" ratio might be 2 and the threshold for "low" ratio 0.6, which are the bounds used in the Current Population Survey (CPS) conducted by U.S. Bureau of the Census (see Kostanich and Dippo 2000, page 10-7). The second criterion is the sample size. A cell whose raw sample count is too small may be collapsed on the grounds that the IAF is unstable. We will refer to a cell as *sparse* if it violates one or the other of the criteria and is collapsed with another cell.

The categories of the variables that define poststrata are usually sorted based on a natural ordering (*e.g.*, age or income categories) or a convenient ordering (*e.g.*, race-ethnicity). Common practice is to collapse a cell with an adjacent one which is similar in characteristics, disregarding different coverage ratios of the individual cells.

Kalton and Flores-Cervantes (2003, page 95) observed that "methods that automatically restrict the range of the adjustments are redistributing the excess adjustments that would otherwise be given to some respondents to other respondents. The appropriateness of this redistribution should be examined." This paper indeed examines its appropriateness and identifies circumstances where the weight redistribution due to collapsing may be quite harmful.

An obvious weakness of popular collapsing strategies is that coverage bias for some groups will be incompletely corrected. For example, suppose that the survey estimate for the number of units in a group is only 1/3 of the census count, so that initial weights would have to be multiplied by 3 to correct for undercoverage. If cell collapsing restricts the

1. Jay J. Kim, National Center for Health Statistics, Centers for Disease Control and Prevention; Jianzhu Li, Joint Program in Survey Methodology, University of Maryland; Richard Valliant, Survey Research Center, University of Michigan.

weight adjustment for units in that group to a factor of 2, then the survey estimate for the number of units in the group will be only 2/3 of the census count. In addition, if cells with much different means are combined, bias can be introduced rather than corrected. The incomplete correction for under-coverage and collapsing of cells with disparate characteristics may lead to bias in totals, means, and other types of estimates.

Table 1 gives some illustrative coverage ratios, *i.e.*, survey estimates prior to poststratification divided by census counts, for the March 2002 U.S. Current Population Survey and the 2003 Behavioral Risk Factors Surveillance Survey (BRFSS) in a set of 44 counties in the southwestern U.S. The survey estimates include a nonresponse adjustment for both CPS and BRFSS. Coverage ratios shown for the subset of demographic groups in the CPS range from 0.70 to 0.93 with Black-only typically being less than for other groups. BRFSS is a telephone survey with low response rates in this set of counties, and the ratios for BRFSS are much smaller than for CPS. There are also substantial differences in coverage ratios for different groups in BRFSS. For example, the ratio for 35 - 44 year old Hispanic males is 0.18 but is 0.37 for Black/Multiracial/Other males in the same age range. If these two groups were collapsed, incomplete coverage would be under-corrected for the Hispanics but over-corrected for the Black/Multiracial/Other group. Another example is the 2003 National Health Interview Survey (NHIS) where American Indians and Asians were collapsed with Whites within age groups. In the cell for ages 25-29, for example, the coverage rates for Whites, American Indians, and Asians were 0.60, 0.44, and 0.31, respectively (Tompkins and Kim 2006).

This paper demonstrates the weaknesses of current cell collapsing procedures and proposes some alternatives. Section 2 discusses the bias of some standard estimators when there is undercoverage. Section 3 introduces two new estimators that retain more of the undercoverage adjustment than the standard method when cells are collapsed. Empirical properties of the standard and alternative methods are investigated through simulation in section 4. We conclude in section 5 with a summary and some possibilities for future research.

2. Some standard estimators

Three standard estimators of the population mean are the Hájek estimator, the poststratified estimator, and the poststratified estimator with initial poststrata collapsed together where necessary. Each of these is defined in detail below. When the sampling frame covers units in the target population at different rates, each estimator can be biased. Kim *et al.* (2005) give some numerical illustrations of the effects of collapsing pairs of cells with different coverage rates.

To derive theory for alternative estimators of means, we model a unit's being covered by the sampling frame and a cell's being sparse or not as random events. Define three indicator variables: $\delta_k = 1$ if unit k is selected for the sample and 0 if not; $c_k = 1$ if unit k is covered by the frame and 0 if not; $d_i = 1$ if poststratum i is classified as sparse in a particular sample and 0 if not $(i = 1, \ldots, I)$. These indicators are assumed to be mutually independent and to have expectations π_k, ϕ_k, and p_i, respectively. Consider a stratified, two-stage probability sample design. A design stratum is denoted by h; s_h is the set of primary sampling units (PSU's) selected from design stratum h; $s_{hj(i)}$ is the set of sample units from sample PSU j in stratum h that are also in poststratum i; U_h is the population of PSU's in stratum h; U_{hj} is the population of units in PSU j within stratum h; and $U_{hj(i)}$ is the population of units in PSU j within stratum h that are in poststratum i. For the analysis in this section, the sample design does not need to be specified in more detail. Note that some summations in sections 2 and 3 of the form $\sum_h \sum_{j \in s_h} \sum_{k \in s_{hj(i)}}$ for units within poststratum i could be simplified to $\sum_{k \in s_{(i)}}$ without loss of generality. We have used the more elaborate notation to make clear how the stages of sampling should be treated.

2.1 Hájek estimator

First, consider the Hájek estimator of a mean, which is

$$\hat{\bar{y}}_\pi = \frac{\left(\sum_{i=1}^{I} \sum_h \sum_{j \in s_h} \sum_{k \in s_{hj(i)}} y_k / \pi_k\right)}{\left(\sum_{i=1}^{I} \sum_h \sum_{j \in s_h} \sum_{k \in s_{hj(i)}} 1/\pi_k\right)} \equiv \hat{T}_\pi / \hat{N}_\pi. \qquad (1)$$

The expectation of $\hat{T}_\pi$ with respect to sampling and the coverage mechanism is $E_c E_\pi(\hat{T}_\pi) = \sum_{i=1}^{I} \sum_h \sum_{j \in U_h} \sum_{k \in U_{hj(i)}} \phi_k y_k \equiv T^c$, where the c superscript denotes "covered". Similarly, the expectation of $\hat{N}_\pi$ is $E_c E_\pi(\hat{N}_\pi) = \sum_{i=1}^{I} \sum_h \sum_{j \in U_h} \sum_{k \in U_{hj(i)}} \phi_k \equiv N^c$. Expanding $\hat{\bar{y}}_\pi$ around (T^c, N^c), its linear approximation is $\hat{\bar{y}}_\pi \doteq T^c / N^c + 1/N^c \times (\hat{T}_\pi - (T^c/N^c)\hat{N}_\pi)$. Next, consider the bias of $\hat{\bar{y}}_\pi$ as an estimator of $\bar{Y} = \sum_{i=1}^{I} T_i / N$ with T_i being the total for the full population of units in poststratum i (not just the covered portion). After some calculation, the bias is

$$\text{bias}(\hat{\bar{y}}_\pi) \doteq \frac{T^c}{N^c} - \frac{1}{N} \sum_{i=1}^{I} T_i = \frac{C_{\phi y}}{\bar{\phi}} \qquad (2)$$

where $\bar{\phi} = \sum \phi_k / N$, $C_{\phi y} = \sum (\phi_k - \bar{\phi})(y_k - \bar{Y}) / N$, $\bar{Y} = T/N$, and $\sum$ denotes the sum over $i, h, j \in U_h$, and $k \in U_{hj(i)}$. Consequently, $\hat{\bar{y}}_\pi$ is biased if there is any correlation between the variable measured, y, and the coverage probability ϕ_k. The bias in (2) is $O(1)$, meaning that it remains important even in large samples.

Table 1
Coverage ratios from the Current Population Survey (CPS) and Behavioral Risk Factors Surveillance Survey (BRFSS). White only and black only mean that only those races were reported by a respondent in the CPS. Residual-only race group includes cases indicating a single race other than white or black, and cases indicating two or more races. Hispanics may be of any race in the BRFSS tabulation

March 2002 Current Population Survey						
Age	White-only		Black-only		Residual-only	
	Male	Female	Male	Female	Male	Female
0 -15	0.93	0.93	0.78	0.79	0.91	0.90
16 - 19	0.90	0.88	0.76	0.81	0.93	0.75
20 - 24	0.79	0.85	0.72	0.77	0.75	0.72
25 - 34	0.83	0.89	0.70	0.76	0.76	0.80
All ages	0.90	0.92	0.78	0.83	0.85	0.84
2003 BRFSS: 44 border counties in Arizona, California, New Mexico, and Texas						
Age	White Non- Hispanic		Black, Multiracial, Other		Hispanic	
	Male	Female	Male	Female	Male	Female
18 - 24	0.19	0.26	0.12	0.24	0.15	0.22
25 - 34	0.20	0.31	0.10	0.16	0.19	0.39
35 - 44	0.28	0.31	0.37	0.25	0.18	0.30
All ages	0.25	0.31	0.25	0.20	0.18	0.31

Sources: Bureau of the Census (2002), Gonzalez, Town, and Kim (2005).

If the coverage probability is the same for every unit in poststratum i, *i.e.*, $\phi_k = \phi(i)$ for any $k \in U_{hj(i)}$, then the approximate bias reduces to $\text{bias}(\hat{\bar{y}}_\pi) \doteq \bar{\phi}^{-1} \sum_i W_i \times (\phi(i) - \bar{\phi})(\bar{Y}_i - \bar{Y})$ where $W_i = N_i/N$ and $\bar{Y}_i = \sum_{h, j \in U_h, k \in U_{hj(i)}} y_k / N_i$. If there is a correlation between the poststratum coverage probabilities and the poststratum means, the Hájek estimator will again be biased, and the bias could be either positive or negative. If the coverage rates or the means are constant across poststrata, *i.e.*, $\phi(i) = \phi_o$ or $\bar{Y}_i = \bar{Y}$, then the Hájek estimator will be unbiased, but poststrata are usually not formed this way. Also, the bias exists even when the appropriate set of poststrata, that subdivide the population into groups with different means, is unknown to the sampler.

2.2 Poststratified mean with no cell collapsing

The poststratified mean is defined as $\hat{\bar{y}}_{\text{PS1}} = 1/N \sum_{i=1}^{I} (N_i/\hat{N}_{\pi i}) \hat{T}_{\pi i}$ where $\hat{T}_{\pi i}$ and $\hat{N}_{\pi i}$ are defined as in (1) but excluding the summation over i. Define $T_i^c = \sum_{h, j \in U_h, k \in U_{hj(i)}} \phi_k y_k$ and $N_i^c = \sum_{h, j \in U_h, k \in U_{hj(i)}} \phi_k$. These are the expected (with respect to the coverage mechanism) total and count of covered units in poststratum i. Expanding $\hat{\bar{y}}_{\text{PS1}}$ around $(T_i^c, N_i^c), i = 1, \ldots, I$, its linear approximation is

$$\hat{\bar{y}}_{\text{PS1}} \doteq \frac{1}{N} \left[\sum_i \frac{N_i}{N_i^c} T_i^c + \sum_i \frac{N_i}{N_i^c} \left(\hat{T}_{\pi i} - \frac{T_i^c}{N_i^c} \hat{N}_{\pi i} \right) \right].$$

The bias of the poststratified estimator is then the first term of this expression minus $\sum_{i=1}^{I} T_i/N$, and after some manipulation, can be written as

$$\text{bias}(\hat{\bar{y}}_{\text{PS1}}) \doteq \sum_{i=1}^{I} W_i \frac{C_{\phi y i}}{\bar{\phi}_i} \qquad (3)$$

where $\bar{\phi}_i = \sum \phi_k / N_i$, $C_{\phi y i} = \sum (\phi_k - \bar{\phi}_i)(y_k - \bar{Y}_i)/N_i$, and $\sum$ denotes the sum over h, $j \in U_h$, and $k \in U_{hj(i)}$. Thus, $\hat{\bar{y}}_{\text{PS1}}$ is biased if there is any correlation between the y variable measured and the coverage probability ϕ_k in any of the poststrata. If the coverage rate is constant at $\phi_k = \phi(i)$ within poststratum i, then the poststratified estimator is approximately unbiased. From (3) it is apparent that poststrata should be formed so that either coverage rates or the y's are homogeneous within each poststratum. This is similar to the recommendations of Eltinge and Yansaneh (1997), Kalton and Maligalig (1991), and Little and Vartivarian (2005) for the formation of nonresponse adjustment cells. In large surveys, the initial set of candidate poststrata is often more extensive than the sample can support. With few exceptions, some of the initial poststrata are collapsed to control weight adjustments. If no collapsing occurs, this is usually because small categories are pre-collapsed based on prior experience in the same or a similar survey. In that sense, PS1 does not really exist in practice. More common is the collapsing approach, PS2, described below.

2.3 Poststratified mean with collapsing

Turning to the poststratified estimator with collapsing, the sparse cells are identified and combined with other cells considered to be their nearest neighbors. This could result in more than one sparse cell being collapsed with a given nonsparse cell. Neighbors can be defined in various ways, *e.g.*, cells with similar estimated coverage rates, $\hat{N}_{\pi i}/N_i$, cells that are adjacent in some substantive sense like nearby income classes, or cells that have similar means on some important survey variables. The general algorithm for collapsing, given an initial set of cells, is:

(1) Compute the collapsing criteria for each cell, *e.g.*, the IAF's, $N_i/\hat{N}_{\pi i}$, and the cell sample sizes;

(2) Identify the sparse cells, *i.e.*, those whose criteria fall outside the bounds for collapsing;

(3) Determine the nearest, nonsparse neighbor of each sparse cell and combine the sparse cell with its neighbor.

The poststratified mean with collapsing is then $\hat{\bar{y}}_{\mathrm{PS2}}=1/N\sum_g(N_g/\hat{N}_{\pi g})\hat{T}_{\pi g}$ where $\hat{T}_{\pi g}=\sum_{i\in A_g}\sum_{h,j\in s_h,k\in s_{hj(i)}} y_k/\pi_k$ and $\hat{N}_{\pi g}$ is defined similarly. Define $T_g^c=\sum_{A_g}T_i^c$ and $N_g^c=\sum_{A_g}N_i^c$ with A_g being the set of poststrata in collapsed group g. Expanding $\hat{\bar{y}}_{\mathrm{PS2}}$ around (T_g^c, N_g^c), for each collapsed group g, gives

$$\hat{\bar{y}}_{\mathrm{PS2}} \doteq \frac{1}{N}\left[\sum_g \frac{N_g}{N_g^c}T_g^c + \sum_g \frac{N_g}{N_g^c}\left(\hat{T}_{\pi g} - \frac{T_g^c}{N_g^c}\hat{N}_{\pi g}\right)\right].$$

It follows that

$$\mathrm{bias}(\hat{\bar{y}}_{\mathrm{PS2}}) \doteq \sum_g W_g \frac{C_{\phi yg}}{\bar{\phi}_g} \tag{4}$$

with

$$W_g=N_g/N,\ \bar{\phi}_g=\sum\phi_k/N_g,\ C_{\phi yg}=\sum(\phi_k-\bar{\phi}_g)(y_k-\bar{Y}_g)/N_g,$$

and the summations in $\bar{\phi}_g$ and $C_{\phi yg}$ are over $i\in A_g, h$, $j\in U_h$, and $k\in U_{hj(i)}$. If ϕ_k is constant within collapsed group g, this estimator is unbiased, but if $\phi_k=\phi(i)$, *i.e.*, the coverage rate is constant within poststratum i but can differ across the poststrata, then the bias becomes

$$\mathrm{bias}(\hat{\bar{y}}_{\mathrm{PS2}}) \doteq \sum_g W_g \frac{C^*_{\phi yg}}{\bar{\phi}_g} \tag{5}$$

with $\bar{\phi}_g=\sum W_{gi}\phi(i)$, $C^*_{\phi yg}=\sum W_{gi}(\phi(i)-\bar{\phi}_g)(\bar{Y}_i-\bar{Y}_g)$, $W_{gi}=N_i/N_g$, and the summations are over $i\in A_g$.

Thus, in the case where $\hat{\bar{y}}_{\mathrm{PS1}}$ will be unbiased, $\hat{\bar{y}}_{\mathrm{PS2}}$ will be biased if poststrata are collapsed together that have different coverage rates and different population means. Since $\bar{\phi}_g$ and $C^*_{\phi yg}$ are both $O(1)$, the bias does not decrease as the sample increases; thus, the bias-squared will eventually be the dominant part of the mean square error. If cells are collapsed, the cells in each group should have the same coverage rates, the same means, or both to avoid bias.

3. Weight restricted estimators

We examine two alternative methods of weight computation when collapsing of poststrata is used, extending work of Kim (2004). The alternatives are designed to be compromises between (a) use of all poststrata and the potential for large weight adjustments and (b) collapsing of poststrata yielding less variable weights but potentially biased estimates. We refer to these as *weight restriction* (WR) methods. The two alternatives presented in this section use cell collapsing but retain a larger share of the weight adjustment for individual cells than does the standard collapsing method.

The first alternative is denoted PS.WR1 and consists of the following algorithm. Denote the maximum allowable weight adjustment by $f_{\max}$ with $f_{\max}>1$.

(1) Execute steps (1) - (3) of the algorithm in section 2.3 for PS2.

(2) Censor any IAF greater than $f_{\max}$ to $f_{\max}$ and adjust each weight in the corresponding initial cell to $\tilde{w}_k=w_k f_{\max}$ with $w_k=1/\pi_k$. For units in cells with IAF $\le f_{\max}$, set $\tilde{w}_k=w_k$.

(3) Compute a collapsing adjustment factor (CAF) for a collapsed group g as

$$\tilde{f}_g = N_g \Big/ \sum_{i\in A_g}\sum_{h,j\in s_h,k\in s_{hj(i)}} \tilde{w}_k .$$

(4) The final adjusted weight is then $\tilde{w}_k\,\tilde{f}_g$ for unit k in group g.

This method will reduce the largest values of the final weight adjustment below the without-collapsing adjustments, $N_i/\hat{N}_{\pi i}$, though there may be one or more groups that have CAF's greater than the $f_{\max}$ cutoff. The control total for group g, N_g, is met in the sense that $\sum_{i\in A_g}\sum_{h,j\in s_h,k\in s_{hj(i)}}\tilde{w}_k\tilde{f}_g=N_g$ but the control totals for the individual cells in A_g are not.

To analyze the properties of PS.WR1, define $A_{g,sp}$ and $A_{g,\overline{sp}}$ to be the sets of sparse and nonsparse poststrata in collapsed group g. PS.WR1 can be expressed as

$$\hat{\bar{y}}_{\mathrm{PS.WR1}} = \frac{1}{N}\sum_g \frac{N_g}{\hat{N}_{g\mathrm{WR1}}}\hat{T}_{g\mathrm{WR1}}$$

where

$$\hat{T}_{g\mathrm{WR1}} = \sum_{i\in A_{g,sp}}\sum_h\sum_{j\in s_h}\sum_{k\in s_{hj(i)}} f_{\max}w_k y_k + \sum_{i\in A_{g,\overline{sp}}}\sum_h\sum_{j\in s_h}\sum_{k\in s_{hj(i)}} w_k y_k$$

and $\hat{N}_{g\mathrm{WR1}}$ has a similar definition with y_k set to 1. The expectation of $\hat{T}_{g\mathrm{WR1}}$ over the coverage, sparseness, and sampling mechanisms is $E_cE_{sp}E_\pi(\hat{T}_{g\mathrm{WR1}})=T_g^c+(f_{\max}-1)\times\tilde{T}_g^c$ where $\tilde{T}_g^c=\sum_{i\in A_g}p_iT_i^c$. Likewise, $E_cE_{sp}E_\pi(\hat{N}_{g\mathrm{WR1}})=N_g^c+(f_{\max}-1)\tilde{N}_g^c$ with $\tilde{N}_g^c=\sum_{i\in A_g}p_iN_i^c$. $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ can be expanded around the expectations of $(\hat{T}_{g\mathrm{WR1}}, \hat{N}_{g\mathrm{WR1}})$. After some manipulation, the approximate bias of $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ becomes

$$\mathrm{bias}(\hat{\bar{y}}_{\mathrm{PS.WR1}}) \doteq \sum_g W_g \frac{C_{\alpha\phi,\,y,\,g}}{(\overline{\alpha\phi})_g} \tag{6}$$

where

$$(\overline{\alpha\phi})_g = \sum \alpha_i \phi_k / N_g, \alpha_i = 1 + (f_{\max} - 1) p_i,$$
$$C_{\alpha\phi, y, g} = \sum (\alpha_i \phi_k - (\overline{\alpha\phi})_g)(y_k - \overline{Y}_g)/N_g,$$

and the summations are over $i \in A_g$, h, $j \in U_h$, and $k \in U_{hj(i)}$. In the case of a common coverage probability in poststratum i, *i.e.*, $\phi_k = \phi(i)$, we have

$$(\overline{\alpha\phi})_g = \overline{\phi}_g + (f_{\max} - 1)(\overline{p\phi})_g$$

and

$$\alpha_i \phi(i) - (\overline{\alpha\phi})g = (\phi(i) - \overline{\phi}_g) + (f_{\max} - 1)(p_i \phi(i) - (\overline{p\phi})_g)$$

where $(\overline{p\phi})_g = \sum_{A_g} W_{gi} p_i \phi(i)$. From this it follows that

$$C_{\alpha\phi, y, g} = C^*_{\phi yg} + (f_{\max} - 1) C_{p\phi, y, g}$$

with

$$C_{p\phi, y, g} = \sum_{A_g} W_{gi} (p_i \phi(i) - (\overline{p\phi})_g)(\overline{Y}_i - \overline{Y}_g).$$

If the cell means $\overline{Y}_i$ are all equal within a collapsed group, then $C^*_{\phi yg} = C_{p\phi, y, g} = 0$ and $\hat{\bar{y}}_{\text{PS.WR1}}$ will be approximately unbiased. In the special case in which coverage is constant in a group, *i.e.*, $\phi(i) = \overline{\phi}_g$, then $C_{\alpha\phi, y, g} = (f_{\max} - 1)$ $\overline{\phi}_g \sum_{i \in A_g} W_{gi}(p_i - \overline{p}_g)(\overline{Y}_i - \overline{Y}_g)$ with $\overline{p}_g = \sum_{A_g} W_{gi} p_i$. Thus, if p_i and $\phi(i)$ are constant within A_g, then $\hat{\bar{y}}_{\text{PS.WR1}}$ will be nearly unbiased even if the cell means $\overline{Y}_i, i \in A_g$, differ. This condition is almost sure to be false as long as one poststratum in a group has a probability of being sparse that is substantially different from the others.

In the case of a common coverage probability in poststratum $i, \phi(i)$, we can also compare the biases of the collapsed cell estimator, $\hat{\bar{y}}_{\text{PS2}}$, with that of $\hat{\bar{y}}_{\text{PS.WR1}}$. Using results in the previous paragraph, the bias in (6) can be expressed as

$$\text{bias}(\hat{\bar{y}}_{\text{PS.WR1}}) \doteq \sum_g W_g \left[\frac{C^*_{\phi yg} / \overline{\phi}_g + (f_{\max} - 1) C_{p\phi, y, g} / \overline{\phi}_g}{1 + (f_{\max} - 1)(\overline{p\phi})_g / \overline{\phi}_g} \right].$$

Since $1 + (f_{\max} - 1)(\overline{p\phi})_g / \overline{\phi}_g \geq 1$, we can use (5) to obtain

$$\begin{aligned} &\left|\text{bias}(\hat{\bar{y}}_{\text{PS.WR1}})\right| \\ &\leq \left| \sum_g W_g \left[C^*_{\phi y, g} / \overline{\phi}_g + (f_{\max} - 1) C_{p\phi, y, g} / \overline{\phi}_g \right] \right| \\ &= \left| \text{bias}(\hat{\bar{y}}_{\text{PS2}}) + (f_{\max} - 1) \sum_g W_g C_{p\phi, y, g} / \overline{\phi}_g \right|. \end{aligned} \qquad (7)$$

If $p_i \phi(i)$ and $\overline{Y}_i$ are uncorrelated, the absolute bias of $\hat{\bar{y}}_{\text{PS.WR1}}$ is less than or equal to that of $\hat{\bar{y}}_{\text{PS2}}$ because $1 + (f_{\max} - 1)(\overline{p\phi})_g / \overline{\phi}_g \geq 1$. When $p_i \phi(i)$ and $\overline{Y}_i$ are correlated, there are two cases to consider: (i) $\text{bias}(\hat{\bar{y}}_{\text{PS2}}) \geq 0$ and (ii) $\text{bias}(\hat{\bar{y}}_{\text{PS2}}) < 0$. In the former, the last line of (7) will be less than or equal to the absolute bias of $\hat{\bar{y}}_{\text{PS2}}$ if

$$\frac{-2\left|\text{bias}(\hat{\bar{y}}_{\text{PS2}})\right|}{f_{\max} - 1} \leq \sum_g W_g C_{p\phi, y, g} / \overline{\phi}_g \leq 0.$$

In case (ii), the requirement is

$$0 \leq \sum_g W_g C_{p\phi, y, g} / \overline{\phi}_g \leq \frac{2\left|\text{bias}(\hat{\bar{y}}_{\text{PS2}})\right|}{f_{\max} - 1}.$$

If the covariance between the probability of being sparse and covered, $p_i \phi(i)$, and the cell means, $\overline{Y}_i$, is small in all groups and the opposite sign of bias $(\hat{\bar{y}}_{\text{PS2}})$, then $\hat{\bar{y}}_{\text{PS.WR1}}$ will be less biased than $\hat{\bar{y}}_{\text{PS2}}$.

The second alternative is denoted PS.WR2 and is intended to exercise more control over the size of the final weight adjustment than does PS.WR1. In PS.WR1 the final adjustment can be larger than $f_{\max}$. PS.WR2 seeks to limit the final adjustment to $f_{\max} = 2$ or some other maximum set in advance. The general idea is to first determine which cells should be collapsed together, as was done for PS.WR1. Then weights in the sparse cells are multiplied by $f_{\max}$. The weights in the non-sparse cell in a collapsed group are then adjusted by a constant factor to bring the estimated population count in the group to the control count. The detailed algorithm for computing weights for PS.WR2 is the following:

(1) Execute steps (1) - (3) of the algorithm in section 2.3 for PS2.

(2) In a group containing at least one non-sparse cell, compute the control total in group g as $N_g = \sum_{i \in A_g} N_i$ and the adjusted weight for all units k in $A_{g, sp}$ as $\tilde{w}_k = w_k f_{\max}$.

(3) Compute the adjusted weight for all units k in $A_{g, \overline{sp}}$ as $\tilde{w}_k = w_k (N_g - \hat{\tilde{N}}_{g, sp}) / \hat{N}_{g, \overline{sp}}$ where $\hat{N}_{g, \overline{sp}} = \sum_{i \in A_{g, \overline{sp}}} \sum_{h, j \in s_h, k \in s_{hj(i)}} w_k$ and $\hat{\tilde{N}}_{g, sp} = \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} \tilde{w}_k$.

(4) The final adjusted weight is then $\tilde{w}_k$ for unit k in group g.

This second weight restricted estimator can be written as $\hat{\bar{y}}_{\text{PS.WR2}} = (1/N) \sum_g \hat{T}_{g\text{WR2}}$ where

$$\begin{aligned} \hat{T}_{gWR2} &= \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} f_{\max} w_k y_k \\ &\quad + \sum_{i \in A_{g, \overline{sp}}} \sum_{h, j \in s_h, k \in s_{hj(i)}} \frac{N_g - f_{\max} \hat{N}_{g, sp}}{\hat{N}_{g, \overline{sp}}} w_k y_k \end{aligned}$$

where $\hat{N}_{g, sp} = \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} w_k$. The expectation of $\hat{T}_{g\text{WR2}}$ over the coverage, sparseness, and sampling mechanisms is

$$E_c E_{sp} E_\pi (\hat{T}_{g\text{WR2}}) = f_{\max} \tilde{T}^c_g + \frac{N_g - f_{\max} \tilde{N}^c_g}{N^c_g - \tilde{N}^c_g} (T^c_g - \tilde{T}^c_g)$$

where $\tilde{N}^c_g = \sum_{i \in A_g} p_i N^c_i$. After some calculation, the approximate bias of $\hat{\bar{y}}_{\text{PS.WR2}}$ can be written as

$$\begin{aligned}\text{bias}(\hat{\bar{y}}_{\text{PS.WR2}}) &\doteq \frac{1}{N} f_{\max} \sum_g \tilde{N}_g^c (\mu_{g,sp}^c - \mu_{g,\overline{sp}}^c) \\ &+ \sum_g W_g \frac{1}{\sum_{A_g} q_i N_i^c} \\ &\times \left[\sum_{A_g} q_i T_i^c - N_g^{-1} \left(\sum_{A_g} q_i N_i^c \right) \left(\sum_{A_g} T_i \right) \right]\end{aligned}$$

where $\mu_{g,sp}^c = \tilde{T}_g^c / \tilde{N}_g^c$ and $\mu_{g,\overline{sp}}^c = \sum_{A_g} q_i T_i^c / \sum_{A_g} q_i N_i^c$. Next, note that in the case of a common coverage probability in cell i, $\phi_k = \phi(i)$,

$$\begin{aligned}&\sum_{A_g} q_i T_i^c - N_g^{-1} \left(\sum_{A_g} q_i N_i^c \right) \left(\sum_{A_g} T_i \right) \\ &= \sum_{A_g} \sum_{h,\, j \in U_h,\, k \in U_{hj(i)}} (q_i \phi_k - (\overline{q\phi})_g)(y_k - \bar{Y}_g) \\ &= \sum_{A_g} N_i (q_i \phi(i) - (\overline{q\phi})_g)(\bar{Y}_i - \bar{Y}_g) \\ &\equiv N_g (C_{\phi y g}^* - C_{p\phi, y, g})\end{aligned}$$

with $q_i = 1 - p_i$, $(\overline{q\phi})_g = \sum q_i \phi_k / N_g$, and $C_{\phi y g}^*, C_{p\phi,y,g}$ were defined previously. Next, use the fact that $\sum_{A_g} q_i N_i^c = N_g^c - \tilde{N}_g^c$ to define $P_g^c = N_g^c / N_g$, the proportion of units covered in group g, and $\tilde{P}_g^c = \tilde{N}_g^c / N_g$, the expected proportion covered in sparse cells in group g. Then, the bias can also be written as

$$\begin{aligned}\text{bias}(\hat{\bar{y}}_{\text{PS.WR2}}) &\doteq \frac{1}{N} f_{\max} \sum_g \tilde{N}_g^c (\mu_{g,sp}^c - \mu_{g,\overline{sp}}^c) \\ &+ \sum_g W_g \frac{C_{\phi y g}^* - C_{p\phi,y,g}}{P_g^c - \tilde{P}_g^c}. \qquad (8)\end{aligned}$$

Judging from (8), $\hat{\bar{y}}_{\text{PS.WR2}}$ will be approximately unbiased if the mean per unit for the units covered by the frame in each collapsed cell is the same in the sparse cells, $\mu_{g,sp}^c$, as in the nonsparse cells, *i.e.*, $\mu_{g,sp}^c = \mu_{g,\overline{sp}}^c$, and the covariances, $C_{\phi y g}^*$ and $C_{p\phi,y,g}$, are both 0. The latter is accomplished by combining cells with the same means, $\bar{Y}_i$. Combining cells with equal coverage rates does not result in $\hat{\bar{y}}_{\text{PS.WR2}}$ being unbiased. This is more restrictive than for $\hat{\bar{y}}_{\text{PS.WR1}}$, which is unbiased if either the coverage rates or the means are the same in all cells in a collapsed group.

4. An empirical investigation

To test some of the ideas presented earlier, we conducted a simulation study of the bias properties of alternative methods of poststratification. We also examined the performance of one variance estimator that is often used in practice.

4.1 Study population

The population used in the simulation was extracted from the 2003 National Health Interview Survey (NHIS) person public-use file. A subset of the NHIS was created with 21,664 persons. These were divided into 25 strata with each having six PSUs. The strata and PSU's are based on those in the NHIS public use file, but sets of three strata were collapsed together to create new design strata for the study population. We used four binary variables (0-1 characteristics) for the simulation, each of which is based on a person's self-report:

Health insurance coverage - whether a person was covered by any type of health insurance;

Physical, mental, or emotional limitation - whether a person was limited in any of these ways;

Medical care delayed - whether a person delayed medical care or not because of cost in last 12 months;

Overnight hospital stay - whether a person stayed overnight in a hospital in last 12 months.

Table 2 shows the percentages of persons with these four characteristics in cells formed by age and sex. These 16 (age × sex) cells are the initial set of poststrata used in estimation. The percentages can vary substantially among the cells, depending on the characteristic. For, example, 18-24 year olds are much more likely to have no health insurance; children under age 5 and the elderly age 65 and over are much more likely to have had a hospital stay. Collapsing cells together that have different means, or proportions in this case, has the potential to introduce bias, as noted earlier.

We also created one artificial binary variable that had a common mean of 0.20 regardless of the unit's poststratum membership. In that case all estimators, including the Hájek estimator, will be unbiased regardless of coverage rates. Also, the conventional thinking that collapsing of cells may reduce variances by smoothing out extreme weight adjustments may hold for this variable.

4.2 Sample design

Two sample PSU's were selected in each stratum with probability proportional to size (PPS) with the size being the count of persons in each PSU. Sampling of PSU's was done with-replacement to simplify variance estimation. If without-replacement sampling had been used, then a more elaborate method of selection and variance estimation would have been needed (see, *e.g.*, Särndal, Swensson, and Wretman 1992, chapter 3). In each sample PSU, 20 persons were selected by simple random sampling without replacement for a total of 1,000 persons in each sample. For each combination of parameters discussed below, 2,000 samples were selected.

Sixteen initial poststrata were used which were the cross of the eight age groups, shown in Table 2, with gender. In

each sample, we computed the estimators of population proportions, described earlier in sections 2-3 - the Hájek estimator, $\hat{\bar{y}}_{\pi}$, the poststratified estimator $\hat{\bar{y}}_{\mathrm{PS1}}$, that uses all 16 poststrata, the poststratified estimator with collapsing of cells, $\hat{\bar{y}}_{\mathrm{PS2}}$, and the two weight-restricted estimators, $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ and $\hat{\bar{y}}_{\mathrm{PS.WR2}}$. The simulation code was written in the R language (R Development Core Team 2005) with extensive use of the R survey package (Lumley 2004, 2005).

4.3 Coverage mechanisms

Five sets of coverage mechanisms, shown in Table 3, were employed to filter the population before the PSU's were sampled. The coverage ratios varied by poststratum and were different for each of the five characteristics for which proportions were estimated. The coverage ratios specific to each of the five characteristics are named C1 through C5 in Table 3. These coverage ratios were artificially created based on the population means for each age and sex group. Poorer coverage was assigned to groups with larger percentages with a characteristic for health insurance coverage and limitations; the opposite was true for delayed medical care and hospital stays. In C5 the coverage ratios are quite variable and are intended to lead to coverage adjustments that vary substantially among the initial set of 16 poststrata. Although the rates in Table 3 are low, they are comparable to or higher than those for BRFSS in Table 1. In applying these rates, we randomly selected a subset of the population to be in the sample frame for each sample that was selected. For example, if the coverage ratio in the poststratum of males younger than 5 years old is 0.9, then 90% of the population in that poststratum was randomly selected to stay in the sampling frame while the rest had a zero probability of being sampled.

4.4 Collapsing rules

We set up situations where the conditions for unbiasedness in sections 2 and 3 can be violated when cells were collapsed in the simulations. Each of the estimators, $\hat{\bar{y}}_{\mathrm{PS2}}$, $\hat{\bar{y}}_{\mathrm{PS.WR1}}$, and $\hat{\bar{y}}_{\mathrm{PS.WR2}}$ involve cell collapses. If the IAF (poststratification factor) in an initial poststratum, $N_i / \hat{N}_{\pi i}$, exceeds the maximum allowable adjustment, $f_{\max}$, or if the cell sample size is less than a minimum, $n_{i,\min}$, we call this poststratum a "sparse" cell and collapse it with a neighboring cell. We used two methods of determining neighbors, designated here as "adjacency" and "close-mean".

In adjacency collapsing, the neighbors of a specific cell are defined as the cells either horizontally or vertically adjacent to it in the age × sex table. For example, in the following, abbreviated table, the neighbors of cell 3 are the shaded cells 2, 4, and 7.

1	5
2	6
3	7
4	8

Table 2 Percentages of persons with four health-related characteristics in groups formed by age and sex

	Population Counts			Percentage of persons with characteristic: Not covered by health insurance			Physical, mental, emotional limitations			Delayed medical care in last 12 months			Hospital stay in last 12 months		
Age	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
< 5	843	795	1,638	10	9	9	4	3	3	3	4	3	17	15	16
5 - 17	2,271	2,082	4,353	13	14	13	10	6	8	4	4	4	2	1	2
18 - 24	998	1,031	2,029	37	31	34	4	4	4	8	11	9	3	14	8
25 - 44	2,971	3,207	6,178	28	23	25	7	7	7	9	10	9	3	10	6
45 - 64	2,421	2,597	5,018	14	14	14	16	19	18	7	11	9	8	10	9
65 - 69	305	384	689	2	1	2	24	29	27	3	8	6	15	14	14
70 - 74	275	344	619	1	1	1	34	32	33	2	5	4	18	15	17
75+	423	717	1,140	1	1	1	41	48	45	2	2	2	22	22	22
Total	10,507	11,157	21,664	18	16	17	12	13	13	6	8	7	7	10	8

Table 3 Coverage ratios used in the simulations

	C1: Not covered by health insurance		C2: Physical, mental, emotional limitations		C3: Delayed medical care in the last 12 months		C4: Hospital stay in last 12 months		C5: Common Mean *Y*	
Age	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
< 5	0.9	0.9	0.9	0.9	0.5	0.5	0.8	0.8	0.9	0.8
5 - 17	0.8	0.8	0.9	0.6	0.5	0.5	0.5	0.5	0.7	0.2
18 - 24	0.5	0.5	0.8	0.8	0.8	0.8	0.5	0.8	0.4	0.4
25 - 44	0.5	0.5	0.8	0.8	0.8	0.8	0.5	0.8	0.6	0.5
45 - 64	0.8	0.8	0.5	0.6	0.8	0.8	0.5	0.8	0.3	0.8
65 - 69	0.9	0.9	0.5	0.6	0.5	0.5	0.8	0.8	0.4	0.4
70 - 74	0.9	0.9	0.5	0.5	0.5	0.5	0.8	0.8	0.2	0.7
75+	0.9	0.9	0.5	0.5	0.5	0.5	0.8	0.8	0.8	0.9

In the adjacency method, a sparse cell was collapsed with the neighbor with the smallest poststratification factor. In close-mean collapsing, a sparse cell was collapsed with the nonsparse cell whose unweighted sample mean was closest to that of the sparse cell.

Two different values of $f_{\max}$ were used in the simulations – $f_{\max} = 2$ and 1.8. Use of $f_{\max} = 1.8$ leads to more collapsing of cells than $f_{\max} = 2$ and exhibits more of the biases (for the characteristics other than the artificial one with a common mean) noted in sections 2 and 3 caused by combining of cells with different means or different coverage rates. The minimum cell size was set to $n_{i,\min} = 25$. Of course, in practice many variations are used to decide which combinations of cells are allowable. We have used just two of the possibilities for illustration in the simulation.

Once all of the sparse cells and their neighbors are identified, the collapsing process proceeds sequentially from cell 1. In a survey with many potential poststrata defined in advance, these procedures might have to be performed iteratively to eliminate all sparse cells. In this simulation, we performed only one round of collapsing.

4.5 Variance estimation

For each of the estimators of a proportion, a linearization variance estimate was calculated. Each of the variance estimators is based on the linear substitute method (*e.g.*, see Särndal *et al.* 1992, chapter 5). The variance estimates for all estimators of proportions were computed using the `svydesign`, `poststratify`, and `svymean` functions in the R `survey` package. The general, theoretical approach is to make a linear approximation for a particular estimator. The linear approximation is rearranged so that the estimator is written as a sum of weighted PSU totals, and the variance estimator for with-replacement PSU sampling is used. The estimators $\hat{\bar{y}}_{\text{PS1}}$, $\hat{\bar{y}}_{\text{PS2}}$, $\hat{\bar{y}}_{\text{PS.WR1}}$, and $\hat{\bar{y}}_{\text{PS.WR2}}$ are treated as standard poststratified estimators for the purposes of variance estimation. For $\hat{\bar{y}}_{\text{PS1}}$, define the following:

$u_k =$ $N_i / \hat{N}_{i\pi}(y_k - \hat{\bar{y}}_i), k \in s_i,$ with $\hat{\bar{y}}_i = \hat{T}_{i\pi}/\hat{N}_{i\pi}$ and s_i being the set of all sample units in poststratum i; u_k is known as a linear substitute;

$\breve{u}_{hj+} = \sum_{i,k \in s_{hj(i)}} w_k u_k$; and

$\bar{\breve{u}}_{h++} = \frac{1}{n_h}\sum_{j \in s_h} \breve{u}_{hj+}$

The variance estimator for $\hat{\bar{y}}_{\text{PS1}}$ is then

$$v(\hat{\bar{y}}_{\text{PS1}}) = \frac{1}{N^2}\sum_h \frac{n_h}{n_h - 1}\sum_{j \in s_h}(\breve{u}_{hj+} - \bar{\breve{u}}_{h++})^2. \qquad (9)$$

For the collapsed stratum estimator, $\hat{\bar{y}}_{\text{PS2}}$, (9) applies with the linear substitute defined as

$$u_k = \frac{N_g}{\hat{N}_{g\pi}}(y_k - \hat{\bar{y}}_g), k \in s_g,$$

with $\hat{\bar{y}}_g = \hat{T}_{g\pi}/\hat{N}_{g\pi}$ and s_g is the set of all sample units in group g.

In the cases of PS.WR1 and PS.WR2, we calculate the final weights as described in section 3 and call the R `poststratify` function. This results in the linear substitute being computed as

$$u_k = \frac{N_g}{\hat{N}_g}(y_k - \hat{\bar{y}}_g) = y_k - \hat{\bar{y}}_g$$

because $\hat{N}_g = \sum_{i \in A_g}\sum_{k \in s_i} \tilde{w}_k = N_g$. The mean $\hat{\bar{y}}_g$ is computed as

$$\begin{aligned}\hat{\bar{y}}_g &= \sum_{i \in A_g}\sum_{k \in s_i} \tilde{w}_k y_k \Big/ \sum_{i \in A_g}\sum_{k \in s_i} \tilde{w}_k \\ &= \sum_{i \in A_g}\sum_{k \in s_i} \tilde{w}_k y_k \Big/ N_g .\end{aligned}$$

The weighted linear substitute is then $u_k = \tilde{w}_k(y_k - \hat{\bar{y}}_g)$ and $\breve{u}_{hj+} = \sum_{i,k \in s_{hj(i)}} \tilde{w}_k u_k$.

In the cases of PS2, PS.WR1, and PS.WR2, these variance estimators do not account for the dynamic nature of cell collapsing which can vary from sample to sample. Consequently, there is a source of variation that is not accounted for, and we can anticipate that the variance estimates will be somewhat too small compared to empirical, simulation variances.

4.6 Simulation results

Tables 4-7 summarize results for coverage correction errors, relative biases of estimated proportions, variances of alternative estimators, and confidence interval coverage using linearization variance estimators. Table 4 shows average absolute coverage correction error, defined as

$$\bar{e} = (DI)^{-1}\sum_{d=1}^{D}\sum_{i=1}^{I}\left|\hat{N}_{di}/N_i - 1\right| \qquad (10)$$

where d is one of the $D = 2{,}000$ samples and $\hat{N}_{di}$ is the estimated number of units in poststratum i based on the final weights for a particular estimator (Hájek, PS2, PS.WR1, or PS.WR2). The value of $\bar{e}$ is 0 for the poststratified estimator with no cell collapsing, PS1, since it corrects coverage error completely in each of the 16 poststrata. To illustrate how the average coverage correction errors can vary, we estimated the proportions for the health insurance and common mean Y variable using the C1 and C5 frame coverage ratios. For most combinations of coverage ratios, collapsing method, and adjustment bound, PS.WR1 more effectively corrects for coverage error than the standard

collapsing estimator, PS2. For example, $\bar{e} = 0.086$ with PS.WR1 for (health insurance, adjacency collapsing, $f_{\max} = 2$) while PS2 has $\bar{e} = 0.120$. In contrast, PS.WR2 is somewhat worse than PS2 in coverage correction.

Table 5 presents the relative biases (relbias), defined as $100\sum_{d=1}^{D}(\hat{\bar{y}}_d - \bar{Y})/\bar{Y}$ where $\hat{\bar{y}}_d$ is one of the estimates of proportion for sample d. The Hájek estimates are badly biased for the first four characteristics since they include no correction for the differential undercoverage among the cells. The relbiases range from -12.1% for limitations to 13.4% for hospital stay. As noted in section 2, the bias can be either positive or negative, depending on the correlation of coverage rates and cell means.

Poststratification with no collapsing of cells (PS1) gives nearly unbiased estimates while the alternatives - PS2, PS.WR1, and PS.WR2 - all introduce a bias when using adjacency collapsing for the first four characteristics. The number of poststrata after collapsing, shown in Table 5, ranges from 6 to 16 when $f_{\max} = 2$ and from 5 to 13 when $f_{\max} = 1.8$. The relative biases of PS2, using adjacency collapsing, range from -4.4% to 6.2% when $f_{\max} = 2$ and from -6.5 to 9.4% when $f_{\max} = 1.8$. With adjacency collapsing, The alternatives, PS.WR1 and PS.WR2, have biases that are intermediate between PS1 (no collapsing) and PS2. PS.WR1, in particular, is reasonably competitive with PS1 in terms of bias with adjacency collapsing. In contrast, close-mean collapsing yields PS2, PS.WR1, and PS.WR2 estimates that are essentially unbiased when $f_{\max} = 2$. With mean collapsing and $f_{\max} = 1.8$, PS2 and PS.WR2 are still somewhat biased, but PS.WR1 compares well with PS1. For the fifth characteristic (Common mean *Y*), all estimators are nearly unbiased, regardless of collapsing method, as expected.

One justification that is conventionally given for collapsing cells is that extreme weights will be reduced and variances of estimates will, in turn, be reduced. Table 6 shows the ratios of the empirical variances of estimated proportions as a proportion of the variance of PS1. The Hájek estimates have variances that are about 12% and 18% smaller than those of PS1 for health insurance and limitations, but are more variable than PS1 for delayed care and hospital stay. These results also make it clear that the variance of a poststratified estimator can be either increased or decreased by collapsing. There are some minor variance gains from using PS2 for some combinations for the first four variables, but with (adjacency, $f_{\max} = 2$) the PS2 variance of hospital stay is 17% larger than that of PS1. With (adjacency, $f_{\max} = 1.8$), PS2 is 23% more variable for hospital stay. PS.WR1 does not have the extreme variances of PS2 in adjacency collapsing; like PS2, PS.WR2 has larger variance for hospital stay in adjacency collapsing. When close-mean, rather than adjacency, collapsing is used, variances of PS2, PS.WR1, and PS.WR2 are much closer to those of PS1. However, for the Common Mean *Y* variable, collapsing always reduces variance. The reductions are almost 20% for adjacency collapsing.

The right-hand section of Table 6 lists the ratios of the empirical mean square errors (MSEs) of estimated proportions as a proportion of the MSE of PS1. With a few exceptions, PS2 is the worst choice of the poststratified estimators for the first four characteristics regardless of the combination of variable, $f_{\max}$, and collapsing method. When $f_{\max} = 1.8$, the choice that leads to more collapsing, the MSEs of PS2 range from 1.8% to 44.2% larger than those of PS1. The MSEs of both PS.WR1 and PS.WR2 are near those of PS1 with the exception of (hospital stay, $f_{\max} = 1.8$, adjacency) where the 6.3% bias of PS.WR2 leads to an MSE 25.6% larger than that of PS1. Close-mean collapsing is preferable to adjacency collapsing, although for the first four characteristics none of the estimators have smaller MSEs than PS1, which does not use collapsing.

The estimators again perform differently for the Common mean *Y* variable. The MSEs of Hájek, PS2, PS.WR1, and PS.WR2 are all less than that of PS1. The Hájek estimator has the smallest MSE, owing to the fact that poststratification is unnecessary to correct bias in estimating the mean.

Table 4 Average absolute coverage correction error as defined in expression (10)

Collapsing Method	Adjustment Bound $f_{\max}$	Hájek	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weight adjustment)
		C1 Coverage ratios for health insurance variable			
Adjacency	2	0.257	0.120	0.086	0.221
Close mean	2	0.257	0.080	0.127	0.281
Adjacency	1.8	0.256	0.150	0.085	0.202
Close mean	1.8	0.256	0.101	0.109	0.258
		C5 Coverage ratios for common mean *Y* variable			
Adjacency	2	0.442	0.326	0.196	0.331
Close mean	2	0.441	0.321	0.203	0.370
Adjacency	1.8	0.442	0.330	0.206	0.376
Close mean	1.8	0.442	0.337	0.214	0.446

Table 5 Relative biases (in percent) of estimated proportions. (Figures for Hájek and PS1 are not affected by collapsing and are repeated in the four sections of the table to facilitate comparisons)

Characteristic	Range of no. of poststrata after collapsing	Hájek	PS1 (no collapsing)	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weight adjustment
			Adjacency collapsing, adjustment bound = 2			
Health insurance	(10, 16)	-11.5	0.1	-4.4	1.0	-1.4
Limitations	(8, 15)	-12.1	-0.3	-2.0	0.1	-1.0
Delayed care	(6, 14)	8.2	-0.2	2.2	-0.6	0.9
Hospital stay	(9, 16)	13.4	0.2	6.2	-0.7	2.8
Common mean *Y*	(5, 11)	0.3	0	0.4	0.4	0.6
			Close-mean collapsing, adjustment bound = 2			
Health insurance	(10, 16)	-11.5	0.1	-0.5	0.5	-0.3
Limitations	(8, 15)	-12.1	-0.3	-1.2	0.2	-1.1
Delayed care	(6, 14)	8.2	-0.2	-0.3	-0.3	-0.2
Hospital stay	(9, 16)	13.4	0.2	0.4	0.1	0.4
Common mean *Y*	(6, 11)	0.3	0	0.2	0.1	0.2
			Adjacency collapsing, adjustment bound = 1.8			
Health insurance	(7, 13)	-11.5	0.1	-6.5	0.7	-3.5
Limitations	(7, 12)	-12.1	-0.3	-3.4	0.3	-2.0
Delayed care	(5, 11)	8.2	-0.2	3.5	-0.4	2.5
Hospital stay	(5, 12)	13.4	0.2	9.4	0.0	6.3
Common mean *Y*	(5, 9)	0.3	0.1	0.5	0.6	0.6
			Close-mean collapsing, adjustment bound = 1.8			
Health insurance	(6, 13)	-11.5	0.1	-1.6	0.3	-1.7
Limitations	(7, 12)	-12.1	-0.3	-2.7	0.9	-2.4
Delayed care	(5, 10)	8.2	-0.2	0.2	-0.3	0.5
Hospital stay	(5, 12)	13.4	0.2	1.5	0.3	2.0
Common mean *Y*	(5, 10)	0.3	0.2	0.3	0.3	0.3

Table 6 Ratio of variances (or MSEs) to the variance (or MSE) of the poststratified estimator (PS1) with no collapsing. (Figures for Hájek are repeated in the four sections of the table to facilitate comparisons)

Characteristic	Ratio of variances to the variance of the poststratified estimator (PS1)				Ratio of MSEs to the MSE of the poststratified estimator (PS1)			
	Hájek	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weight adjustment)	Hájek	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weight adjustment)
				Adjacency collapsing, adjustment bound = 2				
Health insurance	0.877	1.014	1.025	0.991	1.500	1.101	1.018	1.006
Limitations	0.821	0.966	1.035	0.977	1.555	1.008	1.017	0.992
Delayed care	1.099	1.023	1.003	1.000	1.239	1.023	1.000	1.000
Hospital stay	1.290	1.169	1.000	1.073	1.733	1.244	1.000	1.070
Common mean *Y*	0.755	0.805	0.908	0.818	0.752	0.801	0.904	0.826
				Close-mean collapsing, adjustment bound = 2				
Health insurance	0.877	1.013	1.014	1.008	1.500	1.006	1.006	1.006
Limitations	0.821	0.999	1.025	0.994	1.555	1.008	1.017	1.008
Delayed care	1.099	0.997	0.998	1.001	1.239	1.000	1.000	1.000
Hospital stay	1.290	1.011	1.000	1.008	1.733	1.012	1.000	1.012
Common mean *Y*	0.776	0.935	0.974	0.902	0.781	0.933	0.973	0.906
				Adjacency collapsing, adjustment bound = 1.8				
Health insurance	0.877	0.960	1.044	0.976	1.500	1.179	1.024	1.048
Limitations	0.821	0.939	1.032	0.961	1.555	1.034	1.017	1.000
Delayed care	1.099	1.051	0.991	1.032	1.239	1.057	0.989	1.034
Hospital stay	1.290	1.225	1.043	1.201	1.733	1.442	1.023	1.256
Common mean *Y*	0.780	0.815	0.882	0.828	0.779	0.816	0.893	0.829
				Close-mean collapsing, adjustment bound = 1.8				
Health insurance	0.877	1.010	1.006	1.019	1.500	1.018	1.000	1.024
Limitations	0.821	0.983	1.051	0.975	1.555	1.034	1.034	1.017
Delayed care	1.099	1.003	0.995	1.001	1.239	1.000	1.000	1.000
Hospital stay	1.290	1.052	1.001	1.059	1.733	1.035	1.000	1.047
Common mean *Y*	0.771	0.924	0.958	0.876	0.778	0.932	0.959	0.879

Table 7
Coverage rates in percent of 95% confidence intervals computed using *t*-distribution with 25 DF. (Figures for Hájek and PS1 are not affected by collapsing and are repeated in the four sections of the table to facilitate comparisons)

Characteristic	Hájek	PS1 (no collapsing)	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weights adjustment
			Adjacency collapsing, adjustment bound = 2		
Health insurance	75.9	93.8	90.1	94.4	93.6
Limitations	70.9	94.5	93.1	94.5	93.9
Delayed care	92.0	94.0	94.5	94.0	94.6
Hospital stay	82.2	94.5	91.8	94.3	93.9
Common mean *Y*	94.8	93.8	94.6	94.4	94.5
			Close-mean collapsing, adjustment bound = 2		
Health insurance	75.9	93.8	93.7	94.2	94.0
Limitations	70.9	94.5	93.0	94.3	93.5
Delayed care	92.0	94.0	93.6	93.9	93.8
Hospital stay	82.2	94.5	94.3	94.7	94.6
Common mean *Y*	93.7	92.9	92.2	92.5	93.2
			Adjacency collapsing, adjustment bound = 1.8		
Health insurance	75.9	93.8	87.5	94.1	92.4
Limitations	70.9	94.5	92.0	94.2	93.3
Delayed care	92.0	94.0	94.8	94.5	94.5
Hospital stay	82.2	94.5	88.4	94.0	91.1
Common mean *Y*	94.8	94.1	94.3	94.8	94.7
			Close-mean collapsing, adjustment bound = 1.8		
Health insurance	75.9	93.8	92.8	94.3	93.3
Limitations	70.9	94.5	92.3	94.5	93.0
Delayed care	92.0	94.0	93.8	94.0	94.4
Hospital stay	82.2	94.5	93.8	94.6	93.8
Common mean *Y*	94.9	94.5	93.6	93.8	94.8

Table 7 reports the empirical coverages of 95% CI's computed using the estimated proportions and the linearization variance estimator that naturally accompanies each. A *t*-distribution with 25 degrees of freedom is used in all cases. The Hájek coverage rates are extremely poor, as expected, ranging from 70.9% to 92% for the first four characteristics. The poststratified estimators, PS1 and PS.WR1 provide 93.8% to 94.7% coverage, *i.e.*, near the nominal 95%. In contrast, PS2 coverage is somewhat poor for Health insurance and hospitalization, especially for (adjacency, $f_{\max} = 1.8$) where the coverages are 87.5% and 88.4%. Coverage rates for PS.WR2 are slightly less than for PS.WR1 but are reasonably close to nominal. Use of close-mean collapsing generally improves the cases of poor coverage found with adjacency. For Common Mean *Y* coverages are good, ranging from 92.2% to 94.9%.

In summary, the weight-restricted estimators, PS.WR1 and PS.WR2, have some advantage over the standard collapsing estimator, PS2. They are generally less biased and retain more of the undercoverage adjustment than does PS2. However, the most critical element in bias-control is how the cells are collapsed in the first place. Collapsing using nearness of cell means or coverage rates is far more preferable than collapsing using some adjacency criterion based on neither of these. Only when cell means were equal did we observe any gain in MSE from collapsing cells. However, equality of cell means is the exception in practice.

5. Concluding remarks

Designers of surveys of households or establishments often have a lengthy list of poststrata or cells in mind when they develop weighting systems. If the sample size in a poststratum is small or the sample estimate of the population count in a poststratum is much different from an external control count, the poststratum may be collapsed with an adjacent one. The conventional justification for collapsing is that the possibility of creating extreme weights is reduced as are variances of estimates.

However, a poor choice of the method for collapsing has at least two undesirable consequences: (i) deficient frame or sample coverage in some cells is not completely corrected and (ii) estimates from the standard approach to collapsing may be quite biased. The latter problem can result in confidence intervals that cover at much less than the nominal rate. Collapsing leads to bias when coverage rates, cell means, or both are correlated within a collapsed poststratum. The bias can be either positive or negative, depending on the correlation.

Cells should be collapsed based on similarity of coverage rates, population cell means, or both in order to avoid bias. This method of collapsing can be much different from standard procedures that only collapse "adjacent" cells, *e.g.*, by combining contiguous age groups. If the adjacency coincides with cells that have similar coverage rates or

means, no bias results. But, this should be checked rather than assumed.

There are at least two practical issues with collapsing based on cell means. One is that, while the theory directs us to collapse based on population means, in a particular sample we will only have estimates for the population covered by the frame. Coverage may be so deficient that the means of the covered and non-covered parts of the population are substantially different, even within the initial poststrata. This would be a case of "nonignorable non-coverage." If so, poststratification based only on the initial set of cells or combinations of them cannot correct coverage bias. A second practical issue is that data on many items are collected in most surveys. Collapsing based on the cell means for one variable may not work well for other variables. In that case, the compromise, suggested by Little and Vartivarian (2005) for nonresponse adjustment, of collapsing based on some weighted average of the means of an important set of variables should be a good solution.

Extensions of this research would be to examine the performance of the class of calibration estimators in correcting coverage errors. Poststratification is a special case. When categories of qualitative auxiliaries are combined due to small sample sizes or other reasons, the same bias problems we have illustrated here may be introduced in more general calibration estimators. One method of allowing some flexibility to depart from controls while retaining important auxiliaries is already available in Rao and Singh (1997). The effect of their proposals on coverage bias needs to be investigated.

Acknowledgements

The authors are indebted to the Associate Editor and the referees for their careful reviews and constructive comments. This paper reports the general results of research undertaken in part by National Center for Health Statistics (NCHS) staff. The views expressed are attributable to the authors and do not necessarily reflect those of the NCHS. The work of R. Valliant was partially supported by Professional Services Contracts 200-2004-M-09302 and 200-2006-M-17916 between the University of Michigan and the National Center for Health Statistics.

References

Bureau of the Census (2002). *Sources and Accuracy of Estimates for Poverty in the United States*: *2002*. Washington DC.

Eltinge, J., and Yansaneh, I. (1997). Diagnostics for formation of nonresponse adjustment cells with an application to income nonresponse in the U.S Consumer Expenditure Survey. *Survey Methodology*, 23, 37-45.

Fuller, W. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.

Gonzalez, J.F., Town, M. and Kim, J. (2005). Mean square error analysis of health estimates from the Behavioral Risk Factor Surveillance System for counties along the United States/Mexico border region. *Proceedings of the Section on Survey Methods Research*. Alexandria VA: American Statistical Association.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kalton, G., and Maligalig, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, 409-428.

Kim, J.J. (2004). Effect of collapsing rows/columns of weighting matrix on weights. *Proceedings of the Section on Survey Methods Research*, American Statistical Association.

Kim, J.J., Thompson, J.H., Woltman, H.F. and Vajs, S.M. (1982). Empirical results from the 1980 Census Sample Estimation Study. *Proceedings of Section on Survey Research Methods*, American Statistical Association, 170-175.

Kim, J.J., and Tompkins, L. (2007). Comparisons of current and alternative collapsing approaches for improved health estimates. Paper presented at the 11th *Biennial CDC/ASTDR Symposium on Statistical Methods*, in Atlanta, Georgia, April 17-18, 2007.

Kim, J.J., Tompkins, L., Li, J. and Valliant, R. (2005). A simulation study of cell collapsing in poststratification. *Proceedings of the Section on Survey Methods Research*, American Statistical Association.

Kostanich, D., and Dippo, C. (2000). *Current Population Survey*: *Design and Methodology*. Technical paper 63. Washington DC: Department of Commerce.

Lazzeroni, L., and Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of American Statistical Association*, 88, 1001-1012.

Little, R.J.A., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 161-168.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 1-19.

Lumley, T. (2005). Survey: Analysis of complex survey samples. R package version 3.0-1. University of Washington: Seattle.

R Development Core Team (2005). R: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org.

Rao, J.N.K., and Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Tompkins, L., and Kim, J.J. (2006). Evaluation of collapsing criteria in sample weighting. Internal NCHS memorandum.

Tremblay, V. (1986). Practical criteria for definition of weighting classes. *Survey Methodology*, 12, 85-97.