Effects of Collapsing Rows/Columns of Weighting Matrix on Weights

Jay J. Kim

National Center for Health Statistics

In sample weighting operations, we always find cells with differential coverage ratios and low raw sample counts. Some categories are observed to have consistently low coverage ratios. In rare occasions, some categories have very high coverage ratios. When the coverage ratios are too high or too low, or the raw sample counts are too low for some cells in a weighting matrix, they are collapsed. Cell collapsing has been investigated before, but it has never been investigated systematically. In this paper, we will show how much cell weight is artificially transferred from one cell to another when cell collapsing is used. We also show the impact of collapsing on the bias and variance. We introduce new collapsing strategies which correct for the deficiencies of the current approach

Key words: Weighting matrix; cell-collapsing; coverage; coverage ratio; bias; variance.

I. Introduction

Estimation by sampling involves weighting the sample data in the weighting matrix. Before iterational proportional fitting is performed on the weighting matrix, each sample unit is multiplied by the inverse sampling rate and non-interview adjustment factor (there could be more factors such as the inverse sub-sampling rate). These initially weighted sample units are assigned to specific cells of the matrix depending on their characteristics. Before the iterational proportional fitting is performed, ratio and minimum criteria are checked for each row and column of the matrix. If either one of them is not met for a row/column, the row/column is collapsed with another row/column.

The “ratio” of the “ratio criterion" is the inverse coverage ratio. It is the ratio of the 100-percent count (control count) to the initially weighted sample count for the row/column. The ratio greater than one (1) indicates that coverage is low for the group represented by a row/column and vice versa. The threshold for a "high" ratio could be four (4) or two (2). The threshold for a low ratio could be one-fourth (1/4) or one-half (½).

We also collapse a row or a column when its raw sample count is too small. The definition of "small" can depend on the total number of raw sample counts and the size of the weighting matrix or weighting array.

Currently a row/column is collapsed with the next row/column which is similar in characteristics disregarding the coverage ratios of the collapsed rows/columns. As a result, the sample units in the row/column with a lower coverage ratio lose a portion of their due weight to those in the row/column with a higher coverage ratio. If an estimate of a characteristic is based on the sample units in the row/column with the lower coverage ratio, it will be an underestimate. Similarly, if the estimate is based on the units in the other row/column, it would be an overestimate. If the estimate is based on an equal number of sample units from both rows/columns, over- or under-estimation will cancel out each other.

Kalton and Florres-Cervantes mentions 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.” (p.95).Our paper in fact examines its appropriateness and found troubling aspects of it.

Because sample units from many different rows/columns can contribute to the tabulated counts, and there could be multiple weighting areas such as in the decennial census sample weighting, the effect of collapsing rows/columns may not be that conspicuous, unless coverage is consistently low for certain groups. Also if there is only one weighting area - major demographic surveys such as the National Health Interview Survey (NHIS) use only one weighting area for the nation - the effect of collapsing will be shown much more clearly.

II Collapsing rows in a 3x1 weighting matrix

Let N_i be the control count for row i. Similarly we let $w_i^{(0)}$ be the total initially weighted sample count for row i. The weighted total count for row i will be

$$w_i^{(1)} = w_i^{(0)} \frac{N_i}{w_i^{(0)}}$$

We will call $\frac{N_i}{w_i^{(0)}}$ the weighting factor or ratio factor for row i which will be denote by f_i. Note in the

above $w_i^{(1)}$ becomes identical to N_i, the control count.

We will consider two scenarios. They are : i) one row has a perfect ratio but the other a poor ratio; and ii) one row has a low ratio but the other a high ratio. Afterwards, we will generalize the results. We assume the third row has ratio of 1 and meets the minimum criterion.

Case 1. One row has a perfect coverage ratio, but the other a poor ratio

This is the case where the ratio criterion is violated. Suppose

$$\frac{N_1}{w_1^{(0)}} = 4 \text{ and } \frac{N_2}{w_2^{(0)}} = 1 \qquad (1)$$

Row 1 has a low coverage ratio. The ratio of row 1 is only one fourth the ratio of row 2. If rows are not collapsed, the ratio factor for each row is in Table 1.

Table 1 Ratio factor for each row if rows are not collapsed

	Ratio factor
Row 1	4
Row 2	1
Row 3	1

When the two are collapsed, the sample cases in row 1 will lose weight while those in row 2 gain in comparison with the case where each of the rows stands by itself. For convenience we assume $N_1 = N_2$.

The ratio factor for these rows after collapsing in terms of row 1 is

$$\frac{N_1 + N_2}{w_1^{(0)} + w_2^{(0)}} = \frac{2}{5} f_1 \qquad (2)$$

This suggests that the sample cases in row 1 are down weighted by a factor of two fifths (2/5), which will be called the transfer factor.
On the other hand, if the ratio factor is expressed in terms of row 2,

$$\frac{N_1 + N_2}{w_1^{(0)} + w_2^{(0)}} = \frac{8}{5} f_2 \qquad (3)$$

The above equation suggests that the sample cases in row 2 will be weighted up by eight fifths (8/5). For row 2, the transfer factor is 8/5.

Case 2. One row has a low ratio but the other a high ratio

We assume $\frac{N_1}{w_1^{(0)}} = 4$ and $\frac{N_2}{w_2^{(0)}} = \frac{1}{4}$. Then
The ratio factor for each cell, if no rows are collapsed is in Table 2.

Table 2. Ratio factor for each row if rows are not collapsed

	Ratio factor
Row 1	4
Row 2	.25
Row 3	1

The ratio factor for these rows after collapsing in terms of row 1 is

$$\frac{N_1 + N_2}{w_1^{(0)} + w_2^{(0)}} = \frac{2}{17} f_1 \qquad (4)$$

On the other hand, if the ratio factor is expressed in terms of row 2,

$$\frac{N_1 + N_2}{w_1^{(0)} + w_2^{(0)}} = \frac{32}{17} f_2 \qquad (5)$$

The sample cases in row 1 are down weighted by a factor of two seventeenths (2/17) and the sample cases in row 2 over-weighted by a factor of 32 seventeenths (32/17).

In comparison with Table 2, the ratio factor for each row, after collapsing, is changed as follows.

Table 3 Ratio factor for each row if rows 1 and 2 are collapsed

	Transfer factor	Ratio factor
Row 1	0.118	0.471
Row 2	1.882	0.471
Row 3	1.000	1

Case 3. Generalization of all cases

Before we assumed $N_1 = N_2$ and that the ratio factor is 2 for row 1, but it would rarely be the case. Thus we assume $N_2 = c N_1$, where $c > 0$ and that the ratio factor is not 1. As before let $f_1 = \frac{N_1}{w_1^{(0)}}$ and $f_2 = \frac{N_2}{w_2^{(0)}}$.

Then the ratio factor in terms of row 1 is,

$$\frac{N_1 + N_2}{w_1^{(0)} + w_2^{(0)}} = \frac{f_2(1 + c)}{c f_1 + f_2} \; f_1 \qquad (6)$$

Note because of collapsing, the sample units in row 1 now get the transfer factor of $\frac{f_2(1 + c)}{c f_1 + f_2}$ on top of the original ratio factor of f_1. Tables 4 and 5 in the Appendix show transfer factors for various values of c, f_1 and f_2.

In Table 4, we can observe that the row, which has better coverage (lower f), takes away weights from the row which has poorer coverage. For example, when c = 10, f_1 = 2 and f_2 = 3, i.e., $f_2 > f_1$, the transfer factor for row 1 (corresponding to f_1) is 1.43. On the other hand when c= 10, f_1 = 3 and f_2 = 2, i.e., $f_2 < f_1$, the transfer factor for row 1 is 0.69. Note also that c, the size factor of N_2 in comparison with N_1, plays a role on the transfer factor. When f_1 = 2 and f_2 = 3, the transfer factor for row 1 over c = 10, 2 and .5 varies. The factor is largest when c is largest, and it is smallest when c is smallest. That is, when a row has better coverage and is much smaller than the row which has a low coverage rate, the gain is more conspicuous for the former.

Lemma 1. The sum of transfer factors for row 1 and row 2 is 2, if c = 1.
Lemma 2. The sum of transfer factors for row 1 and row 2 is 2 if and only if $f_1 = f_2$ when $c \neq 1$.

The transfer factor in Table 4 stays the same irrespective of individual values of f_1 and f_2, if the ratio of f_1 to f_2 remains the same. The transfer factors can be shown in terms of f_1 / f_2 and c, as shown in Table 5. This table shows the factors in a wider range of c values than Table 4.

II.1 Bias and variance of estimator for Case 1.

Horvitz-Thompson estimator is

$$\hat{X} = \sum_i \sum_k^{n_i} \frac{x_{ik}}{\pi_{ik}}, \qquad (7)$$

where k is the k^{th} sample unit in row i and n_i is the sample count in row i. π_{ik} is interpreted in different ways. Oh and Scheuren, Potter, and Kalton and Maligalig treat it as the selection probability. Singh and Mohl, and Salvucci and Ghosh assume that it is the inclusion probability, or noninterview adjusted selection probability. In this paper, we assume π_{ik} is the inclusion probability which is further adjusted by the ratio factor. What the collapsing does is artificially change the inclusion probability during the process of developing the sample weights. In Case 1, the artificially modified inclusion probability for row 1 is $2.5\,\pi_{1k}$ and the probability in row 2 is $\frac{5}{8}\,\pi_{2k}$.

II.1.1 Conditional bias of the estimator

For deriving the expected values, we assume that the coverage ratio is fixed. The expected value of the estimator is as follows.

$$E\left(\sum_i \sum_k^{N_i} \frac{x_{ik}}{\pi_{ik}}\right) = E\left(\sum_k^{N_1} \frac{x_{1k}}{\pi_{1k}}\right) + E\left(\sum_k^{N_2} \frac{x_{2k}}{\pi_{2k}}\right) + E\left(\sum_k^{N_3} \frac{x_{3k}}{\pi_{3k}}\right)$$

In particular, in case 1, it is

$$\sum_k^{N_1} \frac{2}{5} \frac{x_{1k}}{\pi_{1k}} \pi_{1k} + \sum_k^{N_2} \frac{8}{5} \frac{x_{2k}}{\pi_{2k}} \pi_{2k} + \sum_k^{N_3} \frac{x_{3k}}{\pi_{3k}} \pi_{3k}$$

$$= \frac{2}{5} X_1 + \frac{2}{5} X_1 + X_3 \qquad (8)$$

Where X_i is population total for row i and N_i is the population count in row i. Since population total can be expressed as $X_1 + X_2 + X_3$, the bias of the above estimator is $3(X_1 - X_2)/5$.

In general, the expected value with revised probabilities can be expressed as follows.

$$\sum_k^{N_1} \frac{f_2(1+c)}{c f_1 + f_2} \frac{x_{1k}}{\pi_{1k}} \pi_{1k} + \sum_k^{N_2} \frac{f_1(1+c)}{c f_1 + f_2} \frac{x_{2k}}{\pi_{2k}} \pi_{2k}$$
$$+ \sum_k^{N_3} \frac{x_{3k}}{\pi_{3k}} \pi_{3k}$$
$$= \frac{f_2(1+c)}{c f_1 + f_2} X_1 + \frac{f_1(1+c)}{c f_1 + f_2} X_2 + X_3$$

Thus the bias is

$$\frac{c(f_1 - f_2)}{cf_1 + f_2} X_1 + \frac{f_2 - f_1}{cf_1 + f_2} X_2$$

II.1.2 Conditional variance of the estimator

The variance of Horvitz-Thompson (H-T) estimator in general is,

$$\sum_i \sum_k^{N_i} \frac{(1-\pi_{ik})}{\pi_{ik}} x_{ik}^2 + \sum_i \sum_{k'>k} \frac{(\pi_{ikk'} - \pi_{ik}\pi_{ik'})}{\pi_{ik}\pi_{ik'}} x_{ik}x_{ik'} \quad (9)$$

In case 1, variance due to row 1 is

$$\sum_k^{N_1} \frac{(1-2.5\pi_{1k})}{2.5\pi_{1k}} x_{1k}^2 + \sum_{k'>k}^{N_1} \left(\frac{\pi_{1kk'}}{2.5\pi_{1k}\pi_{1k'}} - 1\right) x_{1k}x_{1k'}$$

$$= \sum_k^{N_1} \frac{(2-5\pi_{1k})}{5\pi_{1k}} x_{1k}^2$$

$$+ \sum_{k'>k}^{N_1} \left(\frac{\pi_{1kk'}}{2.5\pi_{1k}\pi_{1k'}} - 1\right) x_{1k}x_{1k'} \quad (10)$$

Similarly, variance due to row 2 is

$$\sum_k^{N_2} \frac{[1 - (5/8)\pi_{2k}]}{(5/8)\pi_{2k}} x_{2k}^2 + \sum_{k'>k}^{N_2} \left(\frac{\pi_{2kk'}}{5/8\,\pi_{2k}\pi_{2k'}} - 1\right) x_{2k}x_{2k'}$$

$$= \sum_k^{N_2} \frac{(8-5\pi_{2k})}{5\pi_{2k}} x_{2k}^2 +$$

$$\sum_{k'>k}^{N_2} \left(\frac{\pi_{2kk'}}{5/8\,\pi_{2k}\pi_{2k'}} - 1\right) x_{2k}x_{2k'} \quad (11)$$

The difference between this formula and the H-T variance formula is

$$.6\left(\sum_k^{N_2} \frac{x_{2k}^2}{\pi_{2k}} + \sum_{k'>k}^{N_2} \frac{\pi_{2kk'}}{\pi_{2k}\pi_{2k'}} x_{2k}x_{2k'}\right) -$$

$$.6\left(\sum_k^{N_1} \frac{x_{1k}^2}{\pi_{1k}} + \sum_{k'>k}^{N_1} \frac{\pi_{1kk'}}{\pi_{1k}\pi_{1k'}} x_{1k}x_{1k'}\right) \quad (12)$$

By collapsing the rows 1 and 2 and reducing the large weights in row 1, the variance due to row 1 is reduced by the amount of $.6\left(\sum_k^{N_1} \frac{x_{1k}^2}{\pi_{1k}} + \sum_{k'>k}^{N_1} \frac{\pi_{1kk'}}{\pi_{1k}\pi_{1k'}} x_{1k}x_{1k'}\right)$. This is due to the fact that the weights which could have been larger when rows are not collapsed, are scaled down. On the other hand, because of the collapsing, the variance due to row 2 is increased by $.6\left(\sum_k^{N_2} \frac{x_{2k}^2}{\pi_{2k}} + \sum_{k'>k}^{N_2} \frac{\pi_{2kk'}}{\pi_{2k}\pi_{2k'}} x_{2k}x_{2k'}\right)$.

The variance due to row 2 becomes larger after collapsing, because the weight becomes larger.

As was observed with the estimator, again assuming the H-T variance is the same as the population variance, if the characteristic of interest is concentrated in row 1, the variance of the estimate is lowered by the amount given above. Similarly, if a characteristic is about the units solely in row 2, there will be overestimation of variance by the amount above. If the characteristic is from both rows and control counts of the rows are the same, over- and underestimation may balance out depending on x_{ik}.

In general, the variance of the estimator obtained from the collapsed weighting matrix, when row 1 and row 2 are collapsed, is as follows. The variance due to row 1 is

$$\sum_k^{N_1} \frac{f_2(1+c) - (cf_1 + f_2)\pi_{1k}}{(cf_1 + f_2)\,\pi_{1k}} x_{1k}^2 +$$

$$\sum_{k'>k}^{N_1} \frac{f_2(1+c)(cf_1 + f_2)\pi_{1kk'}}{(cf_1 + f_2)^2 \pi_{1k}\pi_{1k'}} x_{1k}x_{1k'} - \sum_{k'>k}^{N_1} x_{1k}x_{1k'} \quad (13)$$

The variance due to row 2 is

$$\sum_k^{N_2} \frac{f_1(1+c) - (cf_1 + f_2)\,\pi_{2k}}{(cf_1 + f_2)\,\pi_{2k}} x_{2k}^2 +$$

$$\sum_{k'>k}^{N_2} \frac{f_1(1+c)(cf_1 + f_2)\pi_{2kk'}}{(cf_1 + f_2)^2 \pi_{2k}\pi_{2k'}} x_{2k}x_{2k'} - \sum_{k'>k}^{N_2} x_{2k}x_{2k'} \quad (14)$$

In short, this example shows the effect of collapsing of rows on bias and variance of an estimate which have the same or different ratios. It has been conjectured that the collapsing will increase the bias but lower the variance. However, this conjecture should be qualified. Only if the sample units contributing to the estimate are concentrated in one of the rows, the conjecture will hold. If about equal number of sample units of two rows contribute to the estimate with similar x_{ik}'s, no such change in bias and variance will occur.

III Collapsing rows in a 3x3 weighting matrix

When the weighting matrix is two dimensional or higher, the impact of collapsing of rows with different

ratios will not be that conspicuous. For dealing with two dimensional weighting matrix, we introduce the following notations.
Let

$w_{ij}^{(0)}$ be the initially weighted sample count in cell {i, j};

$w_{ij}^{(1)} = w_{ij}^{(0)} \frac{N_{i.}}{w_{i.}^{(0)}}$, or the weighted cell count after raking by row in iteration 1,
and

$w_{ij}^{(2)} = w_{ij}^{(1)} \frac{N_{.j}}{w_{.j}^{(1)}}$, or the weighted cell count after raking by column in iteration 1.

We assume the order of raking is by row, then by column. One iteration consists of raking once by row and column, respectively.

$w_{ij}^{(3)}$ and $w_{ij}^{(4)}$ can be defined similarly, but they are for iteration 2. We also define the following weighting factors.

$$f_{i.}^{(1)} = \frac{N_{i.}}{w_{i.}^{(0)}}, \qquad (14)$$

$$f_{.j}^{(2)} = \frac{N_{.j}}{w_{.j}^{(1)}} \qquad (15)$$

Thus if only one iteration is used for raking the sample data, the resulting weight of cell {i,j} will be

$$w_{ij}^{(f)} = w_{ij}^{(0)} f_{i.}^{(1)} f_{.j}^{(2)} \qquad (16)$$

where superscript (f) stands for the final weight.

Suppose we have a weighting matrix whose row structure is the same as the case 1 in section II. That is,

$$\frac{N_{1.}}{w_{1.}^{(0)}} = 4 \text{ and } \frac{N_{2.}}{w_{2.}^{(0)}} = 1 \qquad (17)$$

Then $f_{1.}^{(1)} = 4$ and $f_{2.}^{(1)} = 1$. If we look at collapsing these two rows only (ignoring column collapsing), we can see the same impact of collapsing on bias and variance of the estimates as before. We will look at the two scenarios for columns as for rows.

Case 1. One column has a perfect ratio but the other has a low ratio after raking by row

We assume $f_{.1}^{(1)} = 4$ and $f_{.2}^{(1)} = 1$. We also assume that $N_{1.} = N_{2.}$ and $N_{.1} = N_{.2}$. When rows 1 and 2 are collapsed, and columns 1 and are also collapsed, the transfer factor for each of the cells is in Table 6 on page 8.

Case 2. One column has a low ratio but the other a high ratio

We assume $f_{.1}^{(1)} = 4$ and $f_{.2}^{(1)} = \frac{1}{4}$. We also assume that $N_{1.} = N_{2.}$ and $N_{.1} = N_{.2}$. When columns 1 and 2, in addition to row 1 and 2, are collapsed, the transfer factors will be as shown in Table 7 in the Appendix.

Note all units in cell (1, 1) will get a ratio factor of 0.014. However, if neither row 1 nor column 1 were collapsed, the factor would be 16. Thus, any estimate concentrated in that cell will suffer severe underestimation.

Case 3. Generalization of all cases

Let $N_{2.} = c_1 N_{1.}$ and $N_{.2} = c_2 N_{.1}$. Let i = 1, 2 be defined by i , i', $i \neq i'$, where if i = 2, i' = 1 and vice versa. j and j' are defined similarly for columns. Using equations (6) and (16), the transfer factor for a cell (i, j), i = 1, 2 and j = 1, 2, after one iteration is,

$$\frac{f_{i'.}^{(1)}(1 + c_i)}{c_i f_{i.}^{(1)} + f_{i'.}^{(1)}} \frac{f_{.j'}^{(2)}(1 + c_{i'})}{c_{i'} f_{.j}^{(2)} + f_{.j'}^{(2)}} \qquad (18)$$

Plugging this factor into Horvitz-Thompson estimator, we can obtain the bias and the variance of the estimator after collapsing.

IV An example

Suppose $N_1 = 120$, $W_1 = 80$ (i.e., $f_1 = 1.5$), $N_2 = 1{,}200$, $W_2 = 1{,}200$ ($f_2 = 1$). In this case, $f_1/f_2 = 1.5$ and c = 10. We assume $n_1 = 8$ and $n_2 = 120$.

The weight before collapsing for each sample person in row 1 is 15 (initial weight for a sample person is 10. 10 $f_1 = 15$), but the weight from the collapsed row for each person in original row 1 is 10.3125 [(120 + 1,200)/(80 + 1,200) = 1.03125. Thus 10 times 1.03125 is 10.3125]. Suppose row 1 is for 0-4 year old. The final weighted count for 0-4 year old is 82.5. If collapsing

had not been needed, the count would have been 120, thus 8 sample cases suffer a net loss of 37.5 (37.5/120 = 31.5 or 31.5 percent loss).

The weight before collapsing for a sample person in row 2 is 10, but after collapsing it becomes 10.3125. Supposing row 2 is 5-9 year olds. The total would be 1,237.5 (10.3125x1,200), a net gain of 37.5. For this large category, absorbing additional weight of 37.5 does not affect much. However, for a small category like row 1, the impact is large.

V. Remedies

Several measures are available for avoiding this type of bias. First of all, it is recommended to stay away from the collapsing strategy depending on the contents alone. Rather, it is desirable to develop a statistics-driven strategy. The weighting factor, the coverage ratio and relative control count sizes between the rows/columns involved in collapsing can fill the gap.

For the row/column (row i), which needs to be collapsed with another, identify a row/column which is next to it. Calculate the ratio factor f_i for row i. Multiply every weight in row i by f_i and collapse the two. Perform the raking and obtain the weight. Then multiply every weight in row i by f_i[1]. Note that this strategy can be tweaked. If f_i is too large or too small, we may limit the amount. For example, we can limit f_i to lie between .4 and 1.6.

If the transfer factor is close to 1 for both f_1 and f_2, we combine rows, ignoring the small differences in f_i's. For example, if $.97 \leq f_1/f_2 \leq 1.03$, then collapse rows 1 and 2 as if they have the same f values. Note that when $.97 \leq f_1/f_2 \leq 1.03$, the transfer factor for both f_1 and f_2 will lie between .97 and 1.03.

The alternative is simply to revise the current approach, which is illogical in a certain sense. As mentioned before, the most commonly used ratio criterion for collapsing is 2 and ½ for the upper and lower bounds. According to the current procedure, if a row fails the ratio criterion, the row is collapsed with another row without checking the difference in coverage ratio. That is, if the ratio is 2 for a row, the weights in the row get multiplied by 2. But if the ratio is 3 or 4 for a row, then the row is collapsed and the weights in the row do not get multiplied by 3 or 4, but most likely some number around 1.5, depending on the coverage ratio of the counterpart in the collapsing, which is smaller than 2. In short, until now we have neglected the weights of rows/columns, which have very high or very low coverage ratios, in the pretext of lowering the variance. In order to be consistent in our approaches for collapsing rows/columns, we have to use the ratio 2 for any ratio greater than 2 and ½ for any ratio less than ½. We will show the differences between when f_1 is censored at 2 in Case 1 and not censored at all. Using equation (6), the transfer factor for row 1, when censored, is, $\frac{f_2(1+c)}{2c+f_2}$. Since c = 1 and f_2 = 1, the above transfer factor for row 1 is 2/3 and that for row 2 is 4/3. Note that with the current approach, the weight loss for row 1 is 60 percent, but with the censoring approach at 2, the loss is 33 percent. Similarly, the weight gain with the current approach for row 1 is 60 percent, but the censoring approach reduces the gain by close to half.

Note the above adjustment is much milder than the ones in Table 4. One distinctive feature of this approach is, when both f_1 and f_2 are censored at 2 (which will occur rarely), the transfer factor is 1, thus each sample weight does not get adjusted at all. If anyone wants to avoid this situation, he/she should pick the censoring point which is between f_1 and f_2, which will occur very rarely.

V Concluding Remarks

Thus far observations have been made on the current cell collapsing procedures which redistribute weights without sound statistical grounds. If a small group with a good coverage is collapsed with larger group with poorer coverage, the small group can get large weight gain, but the loss for the large group may not be conspicuous. Thus, the estimate for the small group would be an overestimate. The estimate, based on several groups with different coverages, some of which have weights larger than their own initial weights and the rest have smaller weight canceling out each other, can be a reasonable one. But this type of approach leaves too much up to chance too much. It is desirable to develop some statistically sound strategies, which can guarantee sound estimates. The desirable strategies are based on the weighting factor and transfer factor. Another alternative is simply revising the current approach, giving censored values to the cells, which fail

[1] This criterion was suggested by Dr. Wayne Fuller.

the ratio collapsing criterion. This approach reduces the amount of weight artificially transferred from one cell to another.

VI References

Brackstone, G.J. and Rao, J.N.K. (1976). Raking Ratio Estimators, *Survey Methodology*, Vol. 2, No.1, pp 63-69.

Arora, H.R. and Brackstone, G.J. (1977). An Investigation of the Properties of Raking Ratio Estimators: I. With Simple Random Sampling, *Survey Methodology*, Vol. 3, No.1, pp 62-83.

Brackstone, G.J. and Rao, J.N.K. (1979),. An Investigation of Raking Ratio Estimators, *Sankhya*, C, Vol. 41, pp 97-114.

Deming, W.E. and Stephen, F.F. (1940). On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics*. Vol. XI, No 4, pp 427-444.

Holt, D. and Smith, T.M.F. (1979). Post Stratification, Journal of Royal Statistical Society, Series A, Vol. 142, Part I, pp 33-46.

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, pp 409-428.

Kalton, G. and Florres-Cervantes, I. (2003). Weighting Methods, Journal of Official Statistics, Vol. 19, No. 2, pp 81-97.

Kim, 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, pp 170-175.

Kim, J. (1980). Comparisons of Weighting Methods Based on a Thompson-Willke Test Approach for Population Characteristics. 1980 CENSUS ESTIMATION STUDY, DOCUMENTARY MEMORANDUM NO. 3. Internal Census Bureau memorandum.

Little, R.J.A. (1993). Post-Stratification: a Modeler's Perspective, *Journal of American Statistical Association*, Vol. 88, No. 423, pp 1001-1012.

Oh, H.L. and Scheuren, F. (1983). Weighting Adjustment for Unit Nonresponse. In Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliography, W.G. Meadow, I. Olkin, and D. Rubin (eds). New York: Academic Press.

Potter, F.J. (1990). A Study of Procedures to Identify and Trim Extreme Sampling Weights. Proceedings of Section on Survey Research Methods, American Statistical Association, pp 225-230.

Salvucci, S. and Ghosh, D. (2002). The Fascinating Story of PPS Sampling. American Statistical Association Survey Research Methods Section Newsletter, Issue 14, pp 12-13.

VII. Appendix

Table 4 Transfer factor for row 1 after rows 1 and 2 are collapsed $N_2 = c N_1$

f_1	c = 10.0			c = 2.0			c = .50		
	f_2			f_2			f_2		
	3.0	2.0	1.5	3.0	2.0	1.5	3.0	2.0	1.5
4.0	0.77	0.52	0.40	0.82	0.60	0.47	0.90	0.75	0.64
3.0	1.00	0.69	0.52	1.00	0.75	0.60	1.00	0.86	0.75
2.0	1.43	1.00	0.77	1.29	1.00	0.82	1.13	1.00	0.90
1.5	1.83	1.29	1.00	1.50	1.20	1.00	1.20	1.09	1.00
1.0	2.54	1.83	1.43	1.80	1.50	1.29	1.29	1.20	1.13

Table 5. Transfer factors for row 1 in terms of ratio of coverage ratios when rows 1 and 2 are collapsed

f_1/f_2	c=10	c=5	c=2	c=1	c=.5	c=.2	c=.1

4.0	.268	.286	.333	.400	.500	.666	.786
2.0	.524	.545	.600	.667	.750	.857	.917
1.5	.688	.706	.750	.800	.857	.923	.957
1.25	.815	.828	.857	.889	.923	.960	.978
1.10	.917	.923	.938	.952	.968	.984	.991
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.90	1.10	1.09	1.07	1.05	1.03	1.02	1.01
.75	1.29	1.26	1.20	1.14	1.09	1.04	1.02
.50	1.83	1.71	1.50	1.33	1.20	1.09	1.05
.25	3.14	2.67	2.00	1.60	1.33	1.14	1.07

Table 6. Transfer factor for each cell when both rows 1 and 2 and columns 1 and 2 are collapsed

	Column 1	Column 2	Column 3
Row 1	0.16	0.64	0.4
Row 2	0.64	2.56	1.6
Row 3	0.40	1.60	1

Table 7. Transfer factor for each cell when both rows 1 and 2 and columns 1 and 2 are collapsed

	Column 1	Column 2	Column 3
Row 1	0.014	0.222	0.118
Row 2	0.222	3.542	1.882
Row 3	0.118	1.882	1