# Tips and Tricks for Raking Survey Data (a.k.a. Sample Balancing)

Michael P. Battaglia, David Izrael, David C. Hoaglin, and Martin R. Frankel
Abt Associates, 55 Wheeler Street, Cambridge, MA 02138

**Key Words:** Control totals, convergence, raking margins, weights

## 1. Introduction

A survey sample may cover segments of the target population in proportions that do not match the proportions of those segments in the population itself. The differences may arise, for example, from sampling fluctuations, from nonresponse, or because the sample design was not able to cover the entire population. In such situations one can often improve the relation between the sample and the population by adjusting the sampling weights of the cases in the sample so that the marginal totals of the adjusted weights on specified characteristics agree with the corresponding totals for the population. This operation is known as raking ratio estimation (Kalton 1983), raking, or sample-balancing, and the population totals are usually referred to as control totals. Raking may reduce nonresponse and noncoverage biases, as well as sampling variability. The initial sampling weights in the raking process are often equal to the reciprocal of the probability of selection and may have undergone some adjustments for unit nonresponse and noncoverage. The weights from the raking process are used in estimation and analysis. The adjustment to control totals is sometimes achieved by creating a cross-classification of the categorical control variables (e.g., age categories x gender x race x family-income categories) and then matching the total of the weights in each cell to the control total. This approach, however, can spread the sample thinly over a large number of cells. It also requires control totals for all cells of the cross-classification. Often this is not feasible (e.g., control totals may be available for age x gender x race but not when those cells are subdivided by family income). The use of marginal control totals for single variables (i.e., each margin involves only one control variable) often avoids many of these difficulties. In return, of course, the two-variable (and higher-order) weighted distributions of the sample are not required to mimic those of the population. Raking (or sample-balancing) usually proceeds one variable at a time, applying a proportional adjustment to the weights of the cases that belong to the same category of the control variable. Izrael et al. (2000) introduced a SAS macro for raking (sometimes referred to as the IHB raking macro) that combines simplicity and versatility. More recently, the IHB raking macro has been enhanced to increase its utility and diagnostic capability (Izrael et al. 2004).

## 2. Basic Algorithm

The procedure known as raking adjusts a set of data so that its marginal totals match specified control totals on a specified set of variables. The term “raking” suggests an analogy with the process of smoothing the soil in a garden plot by alternately working it back and forth with a rake in two perpendicular directions.

In a simple 2-variable example the marginal totals in various categories for the two variables are known from the entire population, but the joint distribution of the two variables is known only from a sample. In the cross-classification of the sample, arranged in rows and columns, one might begin with the rows, taking each row in turn and multiplying each entry in the row by the ratio of the population total to the weighted sample total for that category, so that the row totals of the adjusted data agree with the population totals for that variable. The weighted column totals of the adjusted data, however, may not yet agree with the population totals for the column variable. Thus the next step, taking each column in turn, multiplies each entry in the column by the ratio of the population total to the current total for that category. Now the weighted column totals of the adjusted data agree with the population totals for that variable, but the new weighted row totals may no longer match the corresponding population totals. The process continues, alternating between the rows and the columns, and agreement on both rows and columns is usually achieved after a few iterations. The result is a tabulation for the population that reflects the relation of the two variables in the sample.

The above sketch of the raking procedure focuses on the counts in the cells and on the margins of a two-variable cross-classification of the sample. In the applications that we encounter, involving data from complex surveys, it is more common to work with the survey weights of the $n$ individual respondents. Thus, we describe the basic raking algorithm in terms of those individual weights, $w_{i,}\ i = 1, 2, ..., n.$ For an unweighted (i.e., equally weighted) sample, one can simply take the initial weights to be $w_i = 1$ for each $i$.

In a cross-classification that has $J$ rows and $K$ columns, we denote the sum of the $w_i$ in cell $(j, k)$ by $w_{jk}$. To indicate further summation, we replace a subscript by a + sign. Thus, the initial row totals and column totals of the sample weights are $w_{j+}$ and $w_{+k}$ respectively. Analogously, we denote the corresponding population control totals by $T_{j+}$ and $T_{+j}$.

The iterative raking algorithm produces modified weights, whose sums we denote by a suitably subscripted $m$ with a parenthesized superscript for the number of the step. Thus, in the two-variable cross-classification we use $m_{jk}{}^{(1)}$ for the sum of the modified weights in cell $(j,k)$ at the end of step 1. If we begin by matching the control totals for the rows, $T_{j+}$, the initial steps of the algorithm are

$$m_{jk}{}^{(0)} = w_{jk} \qquad (j = 1,...,J;\ k=1,...,K)$$

$$m_{jk}{}^{(1)} = w_{jk}{}^{(0)}\,(T_{j+} / m_{j+}{}^{(0)}) \qquad \text{(for each } k \text{ within each } j\text{)}$$

$$m_{jk}{}^{(2)} = w_{jk}{}^{(1)}\,(T_{+k} / m_{+k}{}^{(1)}) \qquad \text{(for each } j \text{ within each } k\text{)}$$

The adjustment factors, $T_{j+} / m_{j+}^{(0)}$ and $T_{+k} / m_{+k}^{(1)}$, are actually applied to the individual weights, which we could denote by $m_i^{(2)}$, for example. In the iterative process an iteration rakes both rows and columns. Thus, for iteration *s* ( *s* = 0, 1, ...) we may write

$$m_{jk}^{(2s+1)} = m_{jk}^{(2s)} (T_{j+} / m_{j+}^{(2s)})$$
$$m_{jk}^{(2s+2)} = m_{jk}^{(2s+1)} (T_{+k} / m_{+k}^{(2s+1)})$$

In our introductory comments on raking it is important to mention that ideally one should rake on variables that exhibit strong associations with the key survey outcome variables and/or variables that are strongly related to nonresponse or noncoverage. When this is achieved, the mean squared error of the key outcome variables will be reduced. This points to the need to give careful thought to what variables should be used in raking.

## 3. Convergence

Convergence of the raking algorithm has received considerable attention in the statistical literature, especially in the context of iterative proportional fitting for log-linear models, where the number of variables is at least three and the process begins with a different set of initial values in the fitted table (often 1 in each cell). For our purposes it is enough that the iterative raking algorithm (ordinarily) converges, as one would expect from the fact that (in a suitable scale) the fitted cell counts produced by the raking are the weighted-least-squares fit to the observed cell counts in the full cross-classification of the sample by all the raking variables (Deming 1943, Chapter VII).

Convergence may require a large number of iterations. Our experience indicates that, in general, raking on a large number of variables slows the convergence process. However, other factors also affect convergence. One is the number of categories of the raking variables. Convergence will typically be slower for raking on 10 variables each with 5 categories than for 10 variables each with only 2 categories. A second factor is the number of sample cases in each category of the raking variables. Convergence may be slow if any categories contain fewer than 5% of the sample cases. A third factor is the size of the difference between each control total and the corresponding weighted sample total prior to raking. If some differences are large, the number of iterations will typically be higher. One can guard against the possibility of nonconvergence or slow convergence by setting an upper limit on the number of iterations (e.g., 50). Brick et al. (2003) also discuss problems with convergence. They point out that a large number of iterations indicate a raking application that is not “well-behaved” and that problems may exist with the resulting weights – highly variable weights inflate sampling variances and produce unstable domain estimates.

One simple definition of convergence requires that each marginal total of the raked weights be within a specified tolerance of the corresponding control total. As noted above, in practice, when a number of raking variables are involved, one must check for the possibility that the iterations do not converge (e.g., because of sparseness or some other feature in the full cross-classification of the sample). As already noted, one can guard against this possibility by setting an upper limit on the number of iterations. As elsewhere in data analysis, it is sensible to examine the sample (including its joint distribution with respect to all the raking variables) *before* doing any raking. For example, if the sample contains no cases in a category of one of the raking variables, it will be necessary to revise the set of categories and their control totals (say, by combining categories). We recommend that, at a minimum, one should check the unweighted percentage of sample cases in each raking variable category and that the percentage of cases in each control total category also be examined. Small categories in the sample or in the control totals (say under 5%) are potential candidates for collapsing. This will reduce the chance of creating very unequal weights in raking. Category collapsing always needs to be done carefully, and in some instances it may be important to retain a small category in the raking.

## 4. Sources of Control Totals

In our discussion of control totals we are referring to actual totals as opposed to percents. Surveys that use demographic and socioeconomic variables for raking must locate a source for the population control totals. An example of a source of *true* population control totals is the 2000 Census short-form data. The Census long-form variables, the 2000 Census PUMS, the Current Population Survey, Census Bureau population projections, the National Health Interview Survey, and private-sector sources such as Claritas are better viewed as control totals, because they are based either on large samples or on projection methodologies.

For control totals obtained from a sample such as the CPS, the basic idea is that the estimates are subject to much smaller sampling variability and nonresponse bias, and may be subject to much lower noncoverage bias, than a survey sample. For state-specific control totals for, say, persons aged 0-17 years, the CPS estimates will be subject to considerably larger sampling variability; thus they are useful for national control totals, but much less useful for stable state control totals. For projection methods (e.g., age by sex by race mid-year population projections from the U.S. Census Bureau), the basic approach is to project information forward from 2000 for the non-censal years. Clearly, the farther one gets from 2000, the greater the likelihood that the projections will be off. This happened, for example, with the projection of the size of the Hispanic population for the years before the 2000 Census results came out. Eventually, the American Community Survey should provide a new source of information for non-censal years.

It is important to make sure that control totals from different sources all add to the same population total. If not, the raking will not converge. For example, let us assume that one has conducted a survey in the middle of 2003. One is using Census Bureau age, sex, and race projections of the civilian noninstitutionalized population for July 2003. The March 2003 CPS is used to obtain control totals by household income. In this situation one would most likely need to ratio-adjust the CPS income control totals so that they summed to the Census projection control totals for July 2003.

One must also consider how the variables are measured. In a telephone survey one may ask a single question to obtain household income. The source for the control totals may have an income variable that is constructed from a series of questions about income from several sources (wages, cash-assistance programs, interest, dividends, etc.). One needs to consider carefully whether using income as a raking variable makes sense. If the sample is thought to substantially under-represent low-income persons, then raking on income may be preferred. If, on the other hand, there is concern that the survey is measuring income very differently from the source of the control totals, then consideration should be given to raking on a proxy variable such as educational attainment or even a dichotomous poverty-status variable.

Control totals usually do not come with a “missing” category. The same variable in the survey may have a nontrivial percentage of cases that fall in a DK or Refused category. In this situation it may be possible to impute for item nonresponse in the survey before the raking takes place. When imputation is not feasible, the following procedure can be used to adjust the control totals. Run an unweighted frequency distribution on the raking variable in order to determine the percentage of sample cases that have a missing value (e.g., 4.3%). Allocate 4.3% of the control total to a newly created missing category (e.g., 4.3% of 1,500,000 = 64,500). Reapportion the control totals in the other categories so that they add to the reduced control total (1,500,000 – 64,500 = 1,435,500). After raking, the weighted distribution of the sample will agree with the revised control totals and will reflect a 4.3% missing- data rate in weighted frequencies and tabulations.

## 5. Trade-offs Related to the Number of Margins and Number of Categories of Those Margins

Some raking applications use margins for age, sex, and race, because it is relatively easy to obtain control totals for these variables. In other situations (especially in surveys with lower response or important noncoverage issues) one may need to rake on a considerably larger number of variables. This is feasible if control totals can be assembled for these variables. We have seen rakings that used well over ten variables. Raking on many variables will almost always require a large number of iterations. We have also seen rakings that used a smaller number of variables, but with fairly detailed categories. Again, a large number of iterations may be required. In both situations the cross-classification of the raking variables often yields an extremely large number of cells. For example, raking on 12 dichotomous variables yields 4,096 cells. Raking on five variables each containing six categories yields 7,776 cells. Many of these cells will contain no cases in the sample. Such cells, by definition, remain empty after raking. However, the two-variable, three-variable, are higher-order interactions in the sample are maintained in the raking to the marginal control totals. The small cell sizes increase the chance that the raked weights will exhibit considerable variability, because those weights are maintaining sample interactions that are quite unstable.

On top of the challenges of the numbers of variables and categories and the resulting number of underlying cells, if the weighted sample totals before raking differ by a large amount from the marginal control totals, the number of iterations will be even greater. These issues point to the need to closely examine: 1) the variables selected for raking, 2) the number and size of the categories of those raking variables, and 3) the magnitude of differences between the weighted sample totals and the control totals. Ideal variables for raking are those related to the key survey outcome variables and related to nonresponse and/or noncoverage. Variables that do not meet these conditions are candidates for exclusion from raking when a large number of variables are being considered. The categories of each candidate raking variable should be examined to see whether they contain a small proportion of the sample cases (say, under 5%) or whether the control total percentage is small (also, say, under 5%). Such small categories should be considered for collapsing. Sometimes the small categories of a nominal categorical variables can be collapsed into a larger residual category. For ordinal variables, collapsing with an adjacent category is often the best approach. If one or more weighted sample totals differ by a large amount from the corresponding control totals, one should first try to determine the cause of the difference. Is it extreme differential nonresponse, or has the variable in the sample been measured in a very different manner than the corresponding variable used to form the control total? One should consider whether it is appropriate to use such a variable in raking.

## 6. Raking at the State Level in a Large-Scale National Survey

Some large surveys stratify by state and are designed to yield state estimates. The resulting total national sample is usually very large. The survey analysts seek to provide national estimates as well as state estimates. Often one sets up raking control totals at the state level and carries out 51 individual rakings. Let us assume those rakings use variables A, B, and C; but the number of categories of each variable is limited because of the state sample sizes. For example, we often collapse variables A, B, and C differently by state. If variable A were race/ethnicity, we might be able to use Hispanic as a separate race/ethnicity category in California, but not in Vermont due to the small sample size. After the 51 rakings one might compare weighted distribution of variables A, B, and C with national control totals and observe some differences that are caused by the state-level collapsing of categories. If having precise weighted distributions at the national level is important for analytic or “face validity” reasons, one can use the IHB raking macro to do the following:

Set up a single raking that includes margins for State x A, State x B, and State x C (i.e., combine the 51 individual state rakings into a single raking). Then add detailed national margins for variables A, B, and C. Another, similar example would involve adding variable D as a national raking margin because its control total is available only at the national level (e.g., household income). This approach to raking needs to be implemented carefully. Checks should be made for raking variables that contain small sample sizes. The coefficient of variation of the weights prior to raking and after raking should be examined in each state to check for large increases in the variability of the weights. Finally, the raking diagnostics discussed above should be used if convergence problems arise.

### 7. Maintaining Prior Nonresponse and Noncoverage Adjustments in the Final weights

Frankel et al. (2003) have discussed methods based on data on interruptions in telephone service (of a week or longer in the past 12 months) to compensate for the exclusion of persons in nontelephone households in random-digit-dialing surveys. One typically adjusts the base sampling weights of persons with versus without an interruption in telephone service. The resulting interruption-based weight adjusts for the noncoverage of nontelephone households. If one then rakes the sample on the basis of age, sex, and race, the impact of the nontelephone adjustment may be diluted somewhat, even though the interruption-based weight is used as the "input" weight for the raking. In that case it generally makes sense to create weighted control totals (using the interruption-based weight) from the sample for persons residing in households with versus without an interruption in telephone service. These weighted control totals should be ratio-adjusted so that they sum to the age, sex, and race control total. For example, if the age, sex and race margins sum to 180,000,000 persons, then the interruption margin needs to be adjusted so that it also sums to 180,000,000. The raking would use the four variables instead of just three and would ensure that the nontelephone adjustment is fully reflected in the final weights. This would be appropriate where the interruption-in- telephone-service category could be small (e.g., in states where telephone coverage is very high), but one still wants to maintain that small category in the raking.

### 8. Inclusion of Two-Variable Raking Margins

As discussed in Section 2, raking can be viewed as analogous to fitting a main-effects-only model. Because of sample size limitations and/or availability of only one-variable (factor or dimension) control totals, many raking applications follow this approach. In some situations it may be important to fit a two-variable interaction to the data. For example, one is planning to rake on variables A, B, C, and D. However, variable C crossed with variable D is available from the control total source and exhibits a strong interaction (e.g., persons aged 0-17 years are more likely to be Hispanic than persons aged 65+ years). Upon examination of the sample, one determines that the cell counts in the C x D margin are large enough to support fitting a two-variable C x D interaction. In that case one would rake on three margins: A, B, and C x D. It is not necessary to also rake on separate margins for variables C and D. If, however, the C x D raking margin involved collapsing categories of variables C and/or D in order to ensure that cell counts were not too small, one could consider adding one-variable margins to the raking for variables C and D without any collapsing of their categories.

### 9. Forming Control Totals for Quantity Variables

In a specialized raking situation one is planning on raking on some categorical variables. Let us assume that the source of the control totals also has a quantity variable related, to say, the average number of glasses of milk consumed per week. The survey has also measured this same quantity variable; but the survey response rate is, let us assume, only 50%. There is interest in raking to ensure that the weighted total number of glasses of milk consumed per week agrees closely with the control total source. This can be accomplished by dividing the sample into groups (in a simple use, two groups: below and above the median number of glasses of milk consumed from the control total source). Control totals for those groups are used in raking the sample.

### 10. Raking Surveys that Screen for a Specific Target Population

A commonly used survey model for obtaining interviews with a specific target population is to screen a sample of households for the presence of target population members. An example would be children with special health care needs. A roster of children is collected with, say, their age, sex, and race. It is determined whether each child has special health care needs. If the household contains one child with special health care needs, a detailed interview is conducted for that child. If the household has two or more such children, one is selected at random for the detailed interview. Of course, the interview response rate will be less than 100%, because some parents will not agree to do the detailed interview. Let us assume that the survey analysts need to look at prevalence of children with special health care needs, and they will also be analyzing the detailed interview data. In this situation we would calculate the usual base sampling weights, make adjustments for unit nonresponse and possibly make a noncoverage adjustment if warranted. We first obtain control totals for age, sex, and for race in the U.S. population aged 0-17 years. We then rake the entire sample of children in the screened households to those control totals, because that sample is a sample of children aged 0-17 in the U.S. The resulting screener weights can then be used to estimate the prevalence of children with special health care needs in the U.S. That screener weight would typically serve as the input weight in the calculation of weights for the children with completed detailed interviews. As part of that calculation process we seek to weight the detailed interview sample by age, sex, and race. Of course, control totals are unlikely to be available for children with special health care needs. We can, however, use the screener weight and the sample of children with special health care needs identified in the screener survey to form weighted control totals for age, sex, and race and then use them in the raking of the detailed-interview weights. This method ensures that the survey analysts do not ask why the age distribution of children with special health care needs from the screener sample does not agree exactly with the distribution in the detailed interview data. Some caution needs to be exercised in using this approach when there is evidence of false positives from the screener survey.

### 11. Weight Trimming and Raking

Weight truncation and trimming are a separate topic from raking; but they are certainly related, in the sense that weight trimming often takes place at the last step in the calculations, which is often raking. Weight trimming is done in many large-scale surveys (Srinath 2003). Its objective is to reduce the mean squared error of the key outcome estimates. By truncating high weight values one generally lowers sampling variability but may incur some bias. The MSE will be lower if the reduction in variance is large relative to the increase in bias arising from weight truncation. There are no hard-set rules for weight trimming; rather most people use a general set of guidelines. Some rules in common use for the truncation point are: 1) the median weight plus six times the interquartile range (IQR) of the weights, and 2) five times the mean weight. How can weight trimming be incorporated in raking? We have used the IHB SAS macro for weight trimming using the following steps (using the median weight plus six times the IQR):

1. Run the raking to obtain raking weight #1.
2. Examine the distribution of raking weight #1 and calculate cutoff #1 equal to the median plus six times the IQR.
3. Truncate raking weight #1 values above cutoff #1 to cutoff #1 (raking weight #1 values at or below cutoff #1 are not altered).
4. Run the raking using truncated raking weight #1 as the input weight to obtain raking weight #2.
5. Examine the distribution of raking weight #2 and calculate cutoff #2 equal to *one plus* the median plus six times the IQR.
6. Truncate raking weight #2 values above cutoff #2 to cutoff #2 (raking weight #2 values at or below cutoff #2 are not altered).
7. Run the raking using truncated raking weight #2 as the input weight to obtain raking weight #3.
8. Examine the distribution of raking weight #3 and calculate cutoff #3 equal to *one plus* the median plus six times the IQR.
9. Truncate raking weight #3 values above cutoff #3 to cutoff #3 (raking weight #3 values at or below cutoff #3 are not altered).
10. Stop when no weight values exceed the truncation cutoff value.

In steps 5 and 8 we use a cutoff value of *one plus* the median weight plus six times the IQR, because the raking may increase the input weight values of the cases that have been truncated, and thus cause the raking steps to repeat endlessly.

Table 1 shows an example of the use of weight trimming in the IHB SAS macro.

## 12. Conclusions

We have sought to give some background on how raking works and to discuss the convergence process. Our tips and tricks come from extensive use of raking on large-scale surveys. We have also sought to give some warnings of conditions that need to be checked before and after raking. Brick et al. (2003) discuss other examples of issues that one should be aware of when using raking.

The IHB SAS macro discussed in this paper is available for free. If you would like a copy of the macro along with the two SUGI papers, please contact David Izrael at David_Izrael@abtassoc.com.

## References

Bishop, Yvonne M. M., Fienberg, Stephen E., and Holland, Paul W. (1975), *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Brick, J. Michael, Montaquila, Jill, and Roth, Shelly (2003), Identifying Problems with Raking Estimators. *2003 ASA Proceedings* [CD-ROM], Alexandria, VA: American Statistical Association pp. 710-717.

Deming, W. Edwards (1943), *Statistical Adjustment of Data*. New York: Wiley.

Frankel, Martin R., Srinath, K.P., Hoaglin, David C., Battaglia, Michael P., Smith, Philip J., Wright, Robert A., and Khare, Meena (2003), Adjustments for non-telephone bias in random-digit-dialling surveys. *Statistics in Medicine,* Volume 22, pp. 1611-1626.

Izrael, David, Hoaglin, David C., and Battaglia, Michael P. (2000), “A SAS Macro for Balancing a Weighted Sample.” *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference,* Paper 275.

Izrael, D, Hoaglin, D.C., and Battaglia, M.P. (2004), To Rake or Not To Rake Is Not The Question Anymore with the Enhanced Raking Macro. May 2004 SUGI Conference, Montreal, Canada.
Kalton, Graham. (1983), Compensating for Missing Survey Data. Survey Research Center, Institute for Social Research, University of Michigan.

Oh, H. Lock and Scheuren, Fritz (1978), “Some Unresolved Application Issues in Raking Ratio Estimation.” *1978 Proceedings of the Section on Survey Research Methods*, Washington, DC: American Statistical Association, pp. 723-728.

Srinath, K.P. (2003), Internal memorandum on weight trimming, Abt Associates Inc.

Table 1: Example of Weight Trimming During Raking

```
                 OBSERVATIONS IN ORIGINAL DATASET TO BE TRUNCATED
                          CUTOFF: MEDIAN+6*QRANGE
                        CONDITION: MEDIAN+6*QRANGE +1

         weight_to_
   id      truncate       mean      median     qrange     cutoff    condition

   715      477.576     144.250    132.491    51.0592    438.847     439.847
   651      509.018     144.250    132.491    51.0592    438.847     439.847
  1085       690.762     144.250    132.491    51.0592    438.847     439.847
   770      515.720     144.250    132.491    51.0592    438.847     439.847



                 OBSERVATIONS TO BE TRUNCATED AFTER ITERATION = 1
                          CUTOFF: MEDIAN+6*QRANGE
                       CONDITION: MEDIAN+6*QRANGE + 1

         truncated_
   id      weight         mean      median     qrange     cutoff    condition

  1085       451.059     144.250    133.108    51.7302    443.490     444.490


                  OBSERVATIONS TO BE TRUNCATED AFTER ITERATION = 2
                           CUTOFF: MEDIAN+6*QRANGE
                        CONDITION: MEDIAN+6*QRANGE + 1

THERE ARE NO WEIGHTS TO TRUNCATE
RERAKING-TRUNCATION PROCESS CONVERGED IN 2 ITERATIONS WITH CONDITION MEDIAN+6*QRANGE+1
```