IDENTIFYING PROBLEMS WITH RAKING ESTIMATORS

J. Michael Brick, Jill Montaquila, and Shelley Roth
J. Michael Brick, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Calibration estimation, poststratification, auxiliary variables

1. Introduction

Raking was proposed by Deming and Stephan (1940) as a way to ensure consistency between complete counts and sample data from the 1940 U.S. Census of population. Raking is now a widely used procedure that uses auxiliary data from external sources to produce estimates with good statistical properties. The primary use of raking is to adjust the survey estimates for undercoverage and response biases. In addition, raking typically improves the reliability of survey estimates. The effectiveness of raking in reducing the bias and variance of the estimates depends on the relationship between the auxiliary variables used in raking and the survey estimates. When the two are highly correlated, the mean square error of the estimates can be substantially reduced.

Raking, also called iterative proportional fitting, can be described as an algorithm that requires a sequence of adjustments. The survey weights are first adjusted to be consistent with control totals from the marginal distribution of one variable (or dimension). The resulting weights are then adjusted to the control totals for the second marginal distribution. The process continues for all the dimensions. One sequence of adjustments through all the dimensions is called a cycle or iteration. The algorithm iterates until all weighted totals conform to the control totals for all the marginal distributions simultaneously.

In working with a variety of surveys that use raking at Westat, we have encountered some issues and concerns with the procedure. We examine some of these issues in this paper, and provide some recommendations based on our experiences bolstered by theory. We use examples from the 2001 National Household Education Surveys Program (NHES) to illustrate these issues, even though many of the concerns arose in other surveys. NHES is a random-digit-dial (RDD) telephone survey that collects information on the civilian, noninstitutionalized population of the 50 States and the District of Columbia. The NHES surveys are designed to allow repeated measures of various phenomena over time. The objective of NHES:2001 is to make inferences about the entire civilian, noninstitutionalized population for the domains of interest. Although only telephone households are sampled, the estimates are adjusted to totals of persons living in both telephone and nontelephone households derived from the Current Population Survey (CPS). As noted above, one of the main reasons for doing this is to reduce the bias arising from the noncoverage of households without telephones in NHES. NHES:2001 data were used to construct the examples discussed below, but the examples were artificially created to highlight a particular issue. See Hagedorn, et al (2003) for further information about the NHES:2001 Surveys.

In Section 2, we give an overview of raking and poststratification, and present a general model for calibration methods. Properties of raking estimators when there are no empty cells are discussed in Section 3. We describe some potential problems with raking in Section 4. Section 5 contains a summary and some recommendations.

2. Calibration Methods

We begin by describing some theory on the poststratified and raked estimators. Both estimators are in a class of estimators that Deville and Särndal (1992) refer to as calibration estimators. A calibration estimator is one in which the base weights, d_i, the inverse of the probability of selection, are adjusted so that the revised weights, w_i, are close to the original weights but satisfy some constraints. Typically, the constraint is that the sum of the revised weights equal a known population total, $\sum w_i = N$, or more generally $\sum w_i \mathbf{x}_i = \mathbf{X}$, where $\mathbf{x}_i$ is a vector of auxiliary data known for all sampled units and for the entire population. With poststratification, the constraint requires that the revised weights sum to the population total for groups or cells of the population. With raking, the constraints involve summing to the population or control totals for two or more variables at the same time, but the constraints are marginal and do not involve the cells formed by crossing the variables.

Different calibration estimators can be constructed by changing the measure of closeness. For example, raking corresponds to a metric that measures the distance between the revised weight and the original weight (let x denote the ratio) by the function $G(x) = x\log(x) - x + 1$. Deville and Särndal (1992) show the different distance measures have little effect when the sample size is large.

We mentioned some of the reasons for raking in the previous section, but these reasons and others also apply to calibration estimators in general. One reason for using a calibrated estimator is that the constraints force the sum of the weights to equal known population totals for the selected dimensions, and this provides some face validity for the estimates. Since the totals are known from some other source, this is a form of conditioning the estimators that has statistical advantages over unconditional approaches (Holt and Smith, 1979). A second reason for calibrating the estimators is that it adjusts for other sources of error, including sampling, nonresponse, and noncoverage. For example, in many RDD surveys like NHES, adjusting the estimates from a telephone survey to known population counts from the entire population (both telephone and nontelephone households) is the major reason for calibration and is critical to reducing the mean square error of the estimates. The reduction in the mean square error of the estimates is related to the predictive power of the auxiliary variables used in the adjustment. Choosing variables that are highly related to the primary outcome variables of the survey or that are highly related to the propensity to respond or the likelihood of being covered results in the greatest reduction in the mean square error. For example, with a calibrated estimator, the error in estimating a population total that is used as a constraint is zero (it is a known value that does not vary from sample to sample). Similarly, estimates of population totals that are highly correlated with the auxiliary variables have small sampling errors. Below, poststratification and raking are described in more detail.

Consider poststratification to population control totals, where the population count, $N_h \ (h = 1,...H)$, is known for each poststratum or cell *h*. The poststratified estimator of the population total is:

$$\hat{Y}_{ps} = \sum_h N_h \sum_i \left(\frac{d_{hi} y_{hi}}{\sum\limits_i d_{hi}} \right) = \sum_h \sum_i w_{hi} y_{hi} \qquad (1)$$

where d_{hi} is the inverse of the probability of selecting unit *i* in cell *h*, $w_{hi} = \frac{N_h}{\hat{N}_h} d_{hi}$, and $N_h = \sum_{i \in h} d_{hi}$ is the unadjusted survey estimate of the population total in cell *h*.

The poststratified estimator assumes that all the sampled units respond. A generalization used in practice that handles unit nonresponse is the population cell weighting estimator. This estimator is:

$$\hat{Y}_{pc} = \sum_h N_h \left\{ \frac{\sum\limits_{i \in h_r} y_{hi} d_{hi}}{\sum\limits_{i \in h_r} d_{hi}} \right\} = \sum_h \frac{N_h}{\hat{N}_h} \left(\sum_{i \in h_r} y_{hi} d_{hi} \right) \qquad (2)$$

where the sums are over the set of respondents in cell *h*. As a convenience, the population cell weighting estimator is referred to as the poststratified estimator, ignoring the distinction noted above. The difference will be discussed later.

Two issues arise with using the poststratified estimator in practice. The first issue is that the poststratified estimator requires external data on the number of units in the population in each cell (N_h). When several auxiliary variables are used in the adjustment, the counts in each cell of the cross-tabulation of the auxiliary variables are needed, and this level of detail is not always available. Even when the required counts are available, it may not be wise to use the full cross-classification for adjusting the estimator. The poststratified estimator is a ratio estimator, and its denominator is a sample estimate of the number of units in cell *h*. If the sample size in cell *h* is small, then the estimator is biased and the poststratified estimator could be unreliable. Sometimes, the cells are collapsed to avoid this problem.

The population raking estimator is an alternative that may be used when several auxiliary variables are available. Raking is often thought of as a multivariate version of poststratification since the process of raking involves repeated poststratification to multiple dimensions. One virtue of raking is that only the marginal control totals are needed, rather than counts for all the cells in the cross-classification such as would be required with poststratification.

To aid in the discussion, the raking estimator is described in a simple two-variable situation. The extension to more variables is immediate. Suppose there are two auxiliary variables with *H* and *K* classes, respectively. The raking estimator can be written as:

$$\hat{Y}_{pr} = \sum_h \sum_k \tilde{w}_{hk} \left(\sum_{i \in (k,h)_r} y_{hki} d_{hki} \right) \qquad (3)$$

where $\tilde{w}_{hk}$ is the weight formed by raking the weighted count of the number of respondents in cell (*h*,*k*) to the marginal totals as described in the following steps.

- Let $\hat{N}_{hk} = \sum_{i \in (h,k)} d_{(h,k)i}$ be the unadjusted estimate of the population in cell (*h*,*k*).
- Compute weights at each iteration t using the following:

$$\begin{aligned}\tilde{w}_{hk}^{(t)} &= \hat{N}_{hk} && \text{if } t=0\\ &= \frac{\tilde{w}_{hk}^{(t-1)} N_{h+}}{\tilde{w}_{h+}^{(t-1)}} && \text{if } t=\text{odd}\\ &= \frac{\tilde{w}_{hk}^{(t-1)} N_{+k}}{\tilde{w}_{+k}^{(t-1)}} && \text{if } t=\text{even}\end{aligned}$$

- Iterate to convergence (i.e., until the sums of the weights match all the marginal counts within specified tolerance limits).

At each iteration the weights are poststratified twice, first to the row dimension and then to the column dimension. With more than two dimensions, this adjustment is repeated for each dimension. If the raked weights converge, then the order of introducing the dimensions does not matter, and the row and column variables can be interchanged without affecting the estimates. This formulation shows why raking is often viewed as multivariate poststratification.

Another way of writing the raked weight is:

$$\tilde{w}_{hk} = \hat{N}_{hk}\hat{\alpha}_h\hat{\beta}_k \,, \tag{4}$$

where $\hat{\alpha}_h = \lim_{t\to\infty} \prod_{l<t,l\,odd} \frac{N_{h\cdot}}{\tilde{w}_{h\cdot}^{(l)}}$ and $\hat{\beta}_k = \lim_{t\to\infty} \prod_{l<t,l\,even} \frac{N_{\cdot k}}{\tilde{w}_{\cdot k}^{(l)}}$. Using this formulation, the weight can be viewed as being adjusted by a factor for each dimension, α_h is the adjustment for the first dimension (level *h*) and β_k is the adjustment for the second dimension (level *k*). The row factor is the product of all the adjustments that are made to the row across iterations. Similarly, the column factor is the product of all the column adjustments across the iterations.

As an example, suppose that a sample of children between the ages of 5 and 7 enrolled in kindergarten, 1st or 2nd grade are selected in NHES. Control totals are known for both the age and grade margins from an auxiliary source. The control totals are: 3,500, 3,800, and 3,700 for 5, 6, and 7 year olds, respectively; 4,000, 4,000 and 3,000 for kindergarten, 1st grade, and 2nd grade, respectively. Table 1 gives the estimates from the survey before and after raking to the control totals. Using the notation given above, the grade factors are: $\alpha_{kg} = 0.70$, $\alpha_{1st} = 1.05$, and $\alpha_{2nd} = 1.88$, and the age factors are: $\beta_{5yr} = 1.43$, $\beta_{6yr} = 1.08$, and $\beta_{7yr} = 0.68$.

Table 1. Estimates before and after raking by age and grade

Estimates prior to raking				
	Age			
Grade	5	6	7	Total
K	3,200	1,000	100	4,300
1	200	2,500	1,200	3,900
2	0	100	2,200	2,300
Total	3,400	3,600	3,500	10,500
Estimates after raking				
K	3,199	754	47	4,000
1	301	2,843	856	4,000
2	0	203	2,797	3,000
Total	3,500	3,800	3,700	11,000

An important feature of raking that may not be obvious is that the algorithm forces the weights to conform to the marginal totals without perturbing the associations in the unadjusted table (Haberman, 1979). Another way of saying this is that raking retains the cross-product ratios or odds ratios in the observed data, while producing estimates that are consistent with the marginal constraints. For example, in Table 1 notice that all nine cross-product ratios that can be formed are the same for the estimates in tables before and after raking (e.g., the cross-product ratio of the subtable of children aged 5 and 7 and in kindergarten and 1st grade has the value of 192 for both tables).

This feature of raking is implicit when the raked weights are written as a product of row and column factors, such as given by (4). The relationship may be clearer if the full table of survey estimates is written using a log-linear model. For example, when there are three dimensions, the survey estimates are fully determined by the known population total and the model:

$$\begin{aligned}&\ln\left(e_{ijk}\right) = \mu + \alpha_i + \beta_j + \gamma_k + \left(\alpha\beta\right)_{ij} +\\ &\left(\alpha\gamma\right)_{ik} + \left(\beta\gamma\right)_{jk} + \left(\alpha\beta\gamma\right)_{ijk}\end{aligned} \tag{5}$$

where e_{ijk} are the unadjusted estimates in cell *i* of the first dimension, cell *j* of the second dimension, and cell *k* of the third dimension. The first term on the right-hand side of the equation is an intercept term, the next three terms are the main effects corresponding to the specific level for each dimension, the next three terms are the second-order interactions, and the last term is the three-level interaction.

If the survey estimates are poststratified to all the cells in the three dimensional table, then all of the terms in (5) are specified by the control totals, and none of the original structure of the unadjusted estimates are retained. If the survey estimates are raked to the three marginal dimensions instead, then the second- and higher-order terms in (5) are not determined by the control totals. The interaction terms from the unadjusted table are retained in the raked table.

3. Convergence Properties

As noted earlier, raking or iterative proportional fitting is a well-known procedure and many of its properties have been studied in a variety of contexts, but probably most intensively with respect to contingency tables. Ireland and Kullback (1968), and Bishop, Fienberg, and Holland (1972) show that the algorithm converges to a unique solution provided all of the cells in the cross-classified table have positive entries. The proofs require the marginal counts to be consistent. In our terminology, this means the marginal counts all sum to the same total. We expand this definition of consistency in the next section to cover situations in which the same variable is used on more than one dimension, and the control totals do not come from the same source. Convergence is achieved if the raked marginal estimates are equal to the marginal control totals, within a specified tolerance.

While these results are useful in some settings, many of the raking problems in current surveys use auxiliary data to a greater extent than ever before and this often results in raking to tables with zero cells. If some cells have no observations, then the convergence theorems do not apply. Bishop, Fienberg, and Holland (1972) discuss convergence for tables with zero cells, but only for special cases. Fagan and Greenberg (1984) evaluate the convergence when raking with zero cells, but their solution involves iteratively solving linear programming transportation problems just to determine if the problem converges. Given the speed of computing today, it is probably simpler and easier to determine if the algorithm converges by running it and allowing a large number of iterations.

The literature and our own experiences suggest that most well-behaved raking problems converge relatively quickly, even when the tolerance for convergence is relatively small. However, a well-behaved raking problem cannot be defined simply in terms of having no cells with zero counts. Surveys often rake with a large number of dimensions, highly correlated dimensions, or dimensions that have few sample observations for some of the margins. In the next section we consider problems that are not well-behaved.

When the algorithm converges to the specified marginal totals, the procedure has desirable asymptotic properties. For example, Ireland and Kullback (1968) show raking minimizes the discrimination information and produces best asymptotically normal estimates. However, these asymptotic properties may not be very relevant for the finite sample problem, especially if convergence requires a large number of iterations. We believe that a large number of iterations is an indicator of potential problems with the raked weights, frequently signaling that the raking factors may vary substantially. The highly variable weights that are produced by such factors are undesirable because they may inflate the variances of the estimates and result in unstable estimates within domains. Examples of these types of problems are given in the next section.

4. Potential Problems With Raking

In this section, we describe specific situations in which raking could be problematic. Problems associated with inconsistencies in the control totals are discussed in Section 4.1. Section 4.2 examines the problems associated when there are a large number of raking dimensions. In Section 4.3, we discuss raking with dependent dimensions. The potential problems associated with measurement bias are addressed in Section 4.4. Throughout the discussion in Sections 4.1 through 4.4, we consider properties of raking assuming complete response and coverage. The effects of differential nonresponse and coverage on raking are discussed in Section 4.5. Although we discuss each situation individually, it is often the case that these problems are related. (For example, when a large number of dimensions are used, it is often the case that two or more dimensions are correlated.)

With the exception of one operational issue discussed in Section 4.1, we address methodological issues. For example, we do not consider the situation in which the sum of the control totals for one raking dimension does not match the sum of the control totals for another dimension. Although this will result in a failure to converge (provided the difference in the sums of the control totals exceeds the maximum tolerance), it can be easily remedied computationally and is thus strictly an operational and not a methodological concern.

4.1 Inconsistencies in the Control Totals

The one particular operational issue we consider is a particular type of inconsistency in the control totals. We deal with it because it is a relatively common problem and, if overlooked, may have an undesirable effect on the outcome of the raking procedure. In particular, the situation considered here is one in which the same variable is used in more than one raking dimension; e.g., age category by educational attainment is one dimension, and age category by income category is another dimension. If a single external dataset is used as the source for control totals for both of these dimensions, then there should be no inconsistencies in the control totals. However, if different sources are used because the variables are not available from the same external file, inconsistencies may arise. We assume the grand totals from both files are equal so that convergence is possible.

Consider an example in which three age categories (20-39, 40-59, and 60 or older) are used. Suppose that for each age category, three educational attainment categories are used to obtain the first raking dimension, and four income categories are used to obtain the second dimension. Let N_{1ij} denote the control total for age category *i* and educational attainment category *j*, and let N_{2ik} denote the control total for age category *i* and income category *k*. If the totals are from different external sources, for each *i*, $\sum_j N_{1ij} \neq \sum_k N_{2ik}$, generally. Raking will not converge because of this disparity.

However, this operational problem can be easily remedied. The remedy is to compute the control totals for one dimension by applying the proportions (from one source) for that dimension to totals (from the other source) for the other dimension. In the above example, this can be done by applying the proportions from dimension 2 to the totals from dimension 1, as follows:

$$N'_{2ik} = \frac{N_{2ik}}{\sum_k N_{2ik}} \sum_j N_{1ij} \qquad (6)$$

Using this approach, the control totals N_{1ij} and N'_{2ik} are consistent in that $\sum_j N_{1ij} = \sum_k N'_{2ik}$ for all *i*.

Before applying this remedy, it is important to evaluate the quality of the two data sources for the control totals. In this type of calibration problem, the control totals are assumed to be known without error. If the age totals vary substantially across the sources, this premise may not be satisfied and further consideration should be given to the raking.

4.2 Large Number of Dimensions

Now turning to more methodological issues, one situation that could be problematic is when a large number of dimensions are used in raking. This situation may arise in surveys with a large number of auxiliary variables that are associated with nonresponse or undercoverage. It may also arise in longitudinal surveys, due to the wealth of information compiled about respondents at each wave of the survey. Fuller (2002), notes that "If a large number of control variables are used, it may not be possible to construct weights satisfying the calibration constraints and also falling within reasonable bounds." In fact, in these situations, raking may be very slow to converge and, in extreme situations, might not converge at all.

To investigate the behavior of raking with a relatively large number of dimensions, we created an example in which the following dimensions were used for raking:

- Race/ethnicity by income (9 levels);
- Region by urbanicity (8 levels);
- Home tenure by age (18 levels);
- Family type (5 levels); and
- Receipt of food stamps (2 levels).

In all, the cross-classification of these five dimensions results in 12,960 possible cells. The survey dataset used for this example contained only 9,583 cases. Thus, it is obvious that many of the possible cells were empty, and many others had very few cases. In this case, convergence was not achieved in 100 iterations.

Even when the raking does converge, it is important to realize that the associations that are being preserved are based on a sample and are subject to sampling variability, as well as other sources of error. The other sources of error may not be the same for the survey and the source of the control totals. For example, it was noted above that the cross-product ratio of the subtable of children aged 5 and 7 and in kindergarten or 1st grade in Table 1 is 192. This cross-product ratio is computed from the sample cases in the four cells corresponding to a subtable of the full sample. If the sample size in the subtable is small, then the estimated cross-product ratio may not be very stable (even though generally these ratios are stable with moderate cell sizes.)

When the sample sizes in subtables of the raking dimensions are small, then it is possible that the raked weights that preserve the associations based on small samples may have some undesirable features, such as very high variability. This situation can arise in other settings, but it is common when the dimensions are highly correlated, such as with age and grade. We discuss this problem more in the next section.

One remedy is to combine dimensions. Another option is to collapse the levels of the dimensions as needed to retain a large enough sample size in the collapsed cells. Of course, these options may not always be possible or desirable for reasons stated earlier.

4.3 Raking With Dependent Dimensions

With the log-linear model given in (5) in mind, consider the age by grade example discussed in Section 2. Table 2 gives the mean of the weight adjustments in each cell and for both margins. The adjustments can be derived by multiplying the corresponding grade and age factors given in Section 2, or by taking the ratio of the raked estimates to the unadjusted estimates for each cell. Notice that none of the margins for the unadjusted estimates differ by much from the control totals, with the largest mean raking factor of 1.30 applied to the unadjusted 2,300 2nd graders to match the control total of 3,000 2nd graders. However, looking across rows or columns, there is substantial variation in the mean raking factors. For example, in the column corresponding to children who are 6 years old, the adjustment for those in kindergarten is 0.75 and for those in 2nd grade is 2.03. This variation in factors occur even though the overall unadjusted estimate was only 6 percent less than the control total for all 6 year olds.

Table 2. Mean adjustments by cell for the raked survey estimates, by age and grade

	Age			
Grade	5	6	7	Total
K	1.00	0.75	0.47	0.93
1	1.51	1.14	0.71	1.03
2	–	2.03	1.27	1.30
Total	1.03	1.06	1.06	1.05

As noted earlier, with poststratification the survey estimates in the cells defined by the log-linear model in (5) are completely specified by the control totals. For example, when poststratification to cells in a two-way table is used, the model of unadjusted estimates given by $\ln\left(e_{ij}\right)=\mu+\alpha_i+\beta_j+\left(\alpha\beta\right)_{ij}$ is replaced by $\ln\left(p_{ij}\right)=\mu'+\alpha'_i+\beta'_j+\left(\alpha'\beta'\right)_{ij}$, where p_{ij} is the poststratified estimate in cell (*i*,*j*).

With raking, the main effects are fully defined by the control totals, but the raked table retains the higher order interactions from the original table. Consider a table with three dimensions that are raked to each of the three dimensions. The raked table is:

$$\ln\left(r_{ijk}\right)=\mu'+\alpha'_i+\beta'_j+\gamma'_k+\left(\alpha\beta\right)_{ij}+\left(\alpha\gamma\right)_{ik} \\ +\left(\beta\gamma\right)_k+\left(\alpha\beta\gamma\right)_{ijk} \qquad (7)$$

where the higher order interactions are identical to those prior to raking. Only the intercept and main effects are replaced by factors determined by the control totals. In the two-way example with three levels for both dimensions given above, all nine second-order interactions were preserved between the unadjusted and raked tables.

Heuristically, the raking adjustment may result in large variations in the mean cell factors because it is attempting to meet the marginal constraints without perturbing the associations in the original table. If the raking dimensions are based on independent variables (i.e., all the interactions are zero), then raking will converge very quickly. If the dimensions are highly correlated, then more iterations are generally required. With highly correlated dimensions, there is also the possibility that the associations in the unadjusted estimates may not be consistent with control totals and the raking process will never converge. For example, suppose the unadjusted survey estimates fell on the diagonal with 3,400 children who are age 5 and in kindergarten, 3,600 children who are age 6 and in 1st grade, and 3,500 children who are age 7 and in 2nd grade. Assuming the same control totals as given above, the raked estimates never converge.

To summarize, when the raking dimensions are highly correlated, it is not uncommon for the raking process to take a large number of iterations to converge and for the weights to have higher variability than might otherwise be expected. These are in fact the two most common symptoms of problems with raking. In these cases, one possible solution (if feasible) is to combine the two variables to form a single dimension.

In Section 4.1, we discussed an example in which the same variable is used in more than one dimension. Although this will result in correlations between the dimensions, there may be good reasons for doing this. For example, there may be differential coverage among groups formed by combinations of age and education, and further differences in coverage when age is crossed with income. In such situations, it may be desirable to use the same variable in more than one dimension, but it is important to be aware of the potential consequences of using correlated raking dimensions.

4.4 Measurement Bias

When a survey variable is measured with bias the parameter that is estimated is actually a parameter with error, rather than the underlying parameter itself. For example, in (5), with measurement bias in the first dimension, rather than estimating α_i, the survey data may be estimating $\alpha_i^*=\alpha_i+\tau_i$, where τ_i is the measurement bias for level *i* of the first dimension. (We assume here that the control total is the correctly measured variable, but in practice that is not always the case.) If there is measurement bias in a main effect, it is likely that there is also measurement bias in the interaction terms involving the variable that is measured inconsistently.

To investigate the effect of measurement bias on raking, consider an example in which measurement bias was known to exist in one of the dimensions. For this example, four dimensions were used in raking:

- Race/ethnicity by income;
- Region by urbanicity;
- Home tenure by age; and
- An indicator of whether the household received public assistance income.

The public assistance income variable was measured differently in the survey and the source of the control totals. In this case, raking converged rather quickly (in 11 iterations) but there was a considerable amount of variation in the raking adjustment, and even in the mean adjustment for each of the margins for the public assistance income dimension. Because receipt of public assistance income is associated with coverage in a telephone survey, some difference in the mean adjustments for each level of this dimension is to be expected as discussed in the next section. Thus, it may be difficult to identify—and impossible to separate out—the effect of measurement bias. However, if it is known that a variable is measured differently between the survey and the source of the control totals, it is generally good practice not to use that variable in raking (or in other calibration estimators).

In addition to the public assistance example, we encountered another measurement bias example when using both age and grade in raking. Control totals used for raking the person-level weights for NHES are derived from two different CPS files. School enrollment and grade data are available on the October CPS file, but these data are not available on the March CPS file. However, the March CPS file is more contemporaneous with the NHES field period and contains data from the annual demographic survey of the CPS. Thus, the control totals were obtained by applying percentage distributions from the October CPS to an overall estimated total number of children from the March CPS. Marginal grade distributions and marginal age distributions are assumed to remain relatively stable throughout the year. On the other hand, the joint distribution of age and grade changes throughout the year (specifically, over the period from October through March), as children typically remain in the same grade throughout this period while the age distribution within each grade shifts upward. If both age and grade are used in raking—whether in separate dimensions or combined together in a single dimension—this temporal disparity must be addressed. One approach to address and correct for this discrepancy is to "age" or "deage" the sample; i.e., recalculate age to be consistent between the survey and the control totals source.

4.5 Response and Coverage Issues

Earlier it was noted that the poststratified estimator assumed complete response, and the population cell weighting estimator was based on the responding units. More accurately, the poststratified estimator assumes complete response and coverage, while the population cell weighting estimator is based on the responding, covered units. It is assumed that the control totals are based on a source with 100 percent coverage. The same distinction applies with raking, and the raking estimator given by (3) corresponds to the population cell weighting estimator in this regard.

When the data are based on incomplete data (due to either nonresponse or noncoverage), any evaluation of the statistical properties of the survey estimates must be based on an underlying model. Kalton and Maligalig (1991) show that the population cell weighting estimator is unbiased under the response model that assumes all units within a cell have the same probability of being observed (including only response and coverage propensities.) Moreover, they show that if the probability of being observed within a cell is the product of the row and column probabilities, then the population raking estimator is unbiased.

Since most survey estimates are subject to both nonresponse and undercoverage, the implication is that the raked estimates will be biased unless there are no interactions of the auxiliary variables used to create the raking dimensions with the response and coverage rates. In the examples above, this would mean that the probability of being observed may differ by age and grade levels, but the overall probability of being observed must be the product of the age and grade factors to insure unbiasedness under the model.

The practical implications of these results are somewhat limited. First, it is difficult to ascertain whether response rates are products of the row and column effects in most situations. Coverage rates may be estimated in some cases, but the coverage estimates for individual cells are often poorly estimated because of small sample sizes. Second, the data are not expected to conform to the model completely. Often, the bias remaining after adjusting for the row and column effects is smaller than if no adjustments were made. This statement corresponds to the common observation that higher order interactions are generally smaller than lower order effects, but there are exceptions. Third, Little and Wu (1991) show that raking generally works well, even if the model that justifies raking does not hold.

Consequently, we believe it a reasonable and good practice to use raking in these situations, absent reliable data that indicates the model of multiplicative factors for the levels of the dimensions is invalid. If raking is used and the model is inappropriate, then the bias of the raked estimates still will be smaller than bias of the unadjusted estimates in most cases.

5. Summary and Recommendations

Raking is widely used to improve the precision of survey estimates, to force survey totals to match external totals, and to adjust for differential coverage and differential nonresponse. As with other estimation methods, the statistical properties of the raking estimator can be assessed through a model that links the sample to the survey population. Little and Wu (1991) showed that raking generally works well, even if the model that justifies raking does not hold.

In general, raking is an effective estimation approach. However, certain situations may result in problems with raking; these include:

- Inconsistencies in control totals;
- Correlated raking dimensions;
- Sparse tables; and
- Measurement bias.

All of these problems except the first one, which we consider an operational problem, are the result of trying to impose a specific structure on the survey estimates when that structure might not be valid for the survey data. Correlated raking dimensions are problematic when the survey data do not exhibit the same correlation pattern. Sparse tables are problematic when raking tries to preserve the interactions that are based on small sample estimates, while making the weights consistent with the control totals. Measurement bias is problematic because the survey data and control totals are not samples drawn from the same population.

The symptoms of potential problems with raking are slowness or lack of convergence, highly variable overall (mean) adjustments for a given dimension, and highly variable adjustments at the unit level. Since we classify all the problems as a manifestation that the structure of the estimates from the survey is inconsistent with the structure of the values from the population, any of the problems may result in slowness to converge or variability in the weights.

It is imperative that the survey practitioner or analyst who is using raking as an estimation tool include diagnostics in the raking procedure and review the diagnostics for potential problems. The diagnostics should include a count of the number of iterations required for convergence; the mean, minimum, and maximum adjustment factor for each level of each dimension; summary statistics for the unit-level adjustment factor; and summary statistics for the raked weight such as the coefficient of variation of the weights. Such statistics should be compared to similar statistics computed prior to raking.

There are a number of approaches that may be used to remedy problems with raking, depending on the particular situation. In some cases, e.g., when the raking dimensions are correlated, it may be preferable to use poststratification rather than raking, if that approach is feasible. In other cases, problems such as measurement bias might suggest looking for a different source of control totals or dropping a raking dimension. If the problems are due to sparse tables, dropping a raking dimension or collapsing cells (e.g., by combining 2 or more levels of a raking dimension variable) may remedy the problem. Even after the raking procedure has been successfully applied, it may be desirable to trim or truncate the weights; the potential reduction in variance should be weighed against the increase in bias due to trimming or truncation.

6. References

Bishop, Y., Fienberg, S., and Holland, P. (1972). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Deming, W.E., and Stephan, F.F. (1940). On a least-squares adjustment of a sampled frequency table when the expected marginal tables are known. *The Annals of Mathematical Statistics*, 11, pp. 427-444.

Deville, C., and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, pp. 376-382.

Haberman, S. (1979). *Analysis of Qualitative Data*. Academic Press, Inc. New York.

Hagedorn, M., Montaquila, J., Nolin, M.J., Kim, K., Kleiner, B., Waits, T., Chapman, C., and Chandler, K. (2003). *National Household Education Surveys of 2001: Data File User's Manual, Volume I.* U.S. Department of Education, National Center for Education Statistics. NCES 2003-079. Washington, D.C.

Holt, D., and Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society,* Series A, 142, Part 1, pp. 33-46.

Ireland, C.T., and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, pp. 179-188.

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

Little, R.J.A., and Wu, M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86, pp. 87-95.