REDESIGN OF CURRENT POPULATION SURVEY COMPOSITE WEIGHTING

Edwin L. Robison, Martha Duff, and Brandon Schneider, Bureau of Labor Statistics
Harland Shoemaker, Bureau of the Census
Ed Robison, BLS-OEUS-SMD Room 4985, 2 Massachusetts Avenue NE, Washington, DC 20212

Key Words: Raking, Iterative Proportional Fitting, Weighting, Composite Estimation, Current Population Survey

Abstract

Weighting for the monthly Current Population Survey (CPS) includes a process called composite weighting. It is found that for highly correlated items, a lower-variance estimate for the current month can be obtained by also using data from previous months, suitably adjusted with an estimate of change. A composite estimate of this type has been in use for the CPS for decades, but a weighting procedure that could reproduce composite estimates from just a single monthly data file was not implemented until 1998. The procedure now in place successively computes ratio adjustments to weights so that estimates are forced to match three sets of composite estimates of employment, unemployment, and not-in-labor-force 1) by state, 2) by ethnicity x gender x age, and 3) by race x gender x age. A new process to be implemented in January 2003 retains the 3-way iteration, but is designed based on a reevaluation of convergence properties and interaction with a previous step of the estimation. Second-stage weighting is also an iterative proportional fitting (or raking) process with control totals, but most of the second-stage controls are not duplicated by composite weights.

Background

The Current Population Survey is jointly sponsored by the Bureau of Labor Statistics and the Bureau of the Census. The BLS/Census CPS Weighting Group was formed in October 1999 to address weighting issues. CPS weighting includes modules for noninterview adjustment, first-stage ratio adjustment, second-stage ratio adjustment, and benchmarking to composite estimates. Of particular concern were revisions to the weighting needed because of changes in the race and ethnicity questions that will be implemented for the first time in the January 2003 Current Population Survey. We felt this provided an opportunity to look for statistical improvements we could make in the weighting, and at the same time simplify software development and maintenance. Improvement efforts were concentrated on second-stage weighting, since that procedure is most affected by changes in the race and ethnicity questions, and since second-stage adjustment dominates the other adjustments in the CPS weighting process. (See: *Redesign of Current Population Survey Raking to Control Totals* by the same authors as this paper.) Modifying the ensuing composite weighting procedure was also a high priority, since it interacts with the second stage.

In determining what changes to make to composite weighting, a number of factors were considered, including:

- BLS plans for publishing revised race categories at the state and national level (Asian added to core data releases)
- Making control cell definitions more consistent with second-stage weighting and more consistent across the composite weighting steps (state, ethnicity, and race)
- Pre-collapsing small cells to eliminate the need of the current "on-the-fly" collapsing algorithm that produces inconsistent results over time
- Possible changes in race reporting patterns over time
- Simplifying development and maintenance of the weighting software

Description of Current Composite Weighting

The Current Population Survey is a rotating panel survey, obtaining responses from about 50,000 households each month. The primary product of the CPS is labor force data for the Civilian Noninstitutional Population. A given monthly sample is divided into eight panels or rotation groups of households. There is a scheme of panel replacement for the next month where one panel is permanently dropped and replaced by a new panel, and one panel is temporarily dropped for 8 months and replaced by a returning panel. In adjacent months, 6 panels are in common. In a given month one panel each is being interviewed for the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th time ("month-in-sample"). There are known biases in labor force data associated with the month-in-sample.

Current Population Survey composite weighting is preceded by second-stage weighting, an iterative proportional fitting or raking procedure with three steps: a state step, an ethnicity step, and a race step. Each step has population control totals that are estimates of CNP/8, one-eighth of the Civilian Noninstitutional Population. The population controls are essentially 1990 decennial Census estimates, updated to the current time using models and a variety of data sources. (Adjustments for census undercount

are included.) A divisor of 8 is used since each of the eight monthly panels or rotation groups is processed separately. The ethnicity and race steps account for juveniles as well as adults, but the state step is limited to adults 16 years of age and over (CNP16+).

1. The current state step categorizes the CPS sample observations into 51 cells: a single cell for each state and the District of Columbia. Each cell is controlled to its CNP16+ (each panel to CNP16+/8).

2. Following the state step, the current ethnicity step categorizes the same observations into 19 cells: 14 Hispanic gender x age cells and five non-Hispanic age cells, gender combined. Each cell is controlled to its CNP (each panel to CNP/8).
   - Hispanic age categories: 0-5, 6-13, 14, 15, 16-19, 20-29, 30-49, and 50+ (14 and 15 combine gender)
   - Non-Hispanic age categories: 0-5, 6-13, 14, 15, and 16+

3. The last step of the iterative procedure, the race step, categorizes the observations into 118 cells: 66 White gender x age cells, 42 Black gender x age cells, and 10 Other gender x age cells. Each cell is controlled to its CNP (each panel to CNP/8).
   - White age categories: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10-11, 12-13, 14, 15, 16, 17, 18, 19, 20-24, 25-26, 27-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-62, 63-64, 65-67, 68-69, 70-74, and 75+
   - Black age categories: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14, 15, 16-17, 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, and 65+
   - Other age categories: 0-5, 6-13, 14, 15, 16-44, and 45+ (14 and 15 combine gender)

The term "Other" for race is a residual category that includes American Indian, Aleut, Eskimo, Asian, and Pacific Islander. We cannot now distinguish Asians from Native Hawaiians and Other Pacific Islanders, but this will be possible when new race data collection procedures are implemented January 2003.

The steps are iterated separately – six times – for each of the eight CPS panels. For each cell k of each step, a simple adjustment is computed for each of the eight panels using adjusted observation weights w_{kij} from the previous step. The multiplicative panel adjustments are then applied to the weight of every observation in the cell. The adjusted observation weights w_{kij}' are used in the next step of the iterative procedure.

$$ADJ_{ki} = (CNPk/8)/ \sum_j w_{kij}$$
$$w_{kij}' = ADJ_{ki} * w_{kij}$$

Using current second-stage weights, composite estimates are made for employment and unemployment. These composite estimates are then used as controls in a composite weighting procedure. Both employment and unemployment are controlled in each defined cell, unless the cell is collapsed. Not-in-labor-force is controlled as a residual, subtracting composite estimates of employment and unemployment from the CNP16+ for a cell.

- Month t composite estimator for employed:
  $Y^c_t = .6Y^{ss}_t + .4(Y^c_{t-1} + \Delta_t) + .3\beta_t$
- Month t composite estimator for unemployed:
  $Y^c_t = .3Y^{ss}_t + .7(Y^c_{t-1} + \Delta_t) + .4\beta_t$

The formulas are basically weighted averages of this month's simple weighted estimate using second-stage weights (Y^{ss}_t) and the composite estimate Y^c_{t-1} from last month. The composite estimate from last month is updated to the current month by an estimate of change Δ_t developed from the six continuing panels that are in common between last month and this month (specified by month-in-sample 2,3,4,6,7,8 for month t in the next formula). Usually β_t is characterized as an adjustment for month-in-sample bias. The sum of second-stage weights is $x_{t,i}$ for month t, month-in-sample i.

- $\Delta_t = (4/3)\sum(x_{t,i} - x_{t-1,i-1})$ sum over i=2,3,4,6,7,8
- $\beta_t = x_{t,1} + x_{t,5} - (1/3)\sum(x_{t,i})$ sum over i=2,3,4,6,7,8

Compared to estimates made using second-stage weights, composite estimates have lower variances for month-to-month changes in unemployment levels and employment levels. Prior to 1998, composite estimation was done only at the macro level and reproducing official CPS estimates required using several monthly data files. A composite weighting procedure was introduced January 1998, and the resulting composite weights on a monthly data file can be used to reproduce a core of "important" composite estimates.

CPS composite weighting applies only to adults (16+). Like seond-stage weighting, it is an iterative procedure with three steps: a state step, an ethnicity step, and a race step. All eight rotation groups are combined for composite weighting.

1. State Step -- Observations are categorized into 51 cells: a single cell for each state and the District of Columbia. Each cell is controlled to composite estimates of employment and unemployment and a residual for NILF.

2. Ethnicity Step -- Observations are categorized into 9 cells: 8 Hispanic gender x age cells and 1 non-Hispanic cell, gender combined. Each cell is controlled to composite estimates of employment and unemployment and a residual for NILF.
   - Hispanic age categories: 16-19, 20-29, 30-49, and 50+
   - Non-Hispanic age category: 16+ (combine gender)

3. Race Step -- Observations are categorized into 66 gender x age cells: 38 White cells, 24 Black cells, and 4 Other cells. Each cell is controlled to composite estimates of employment and unemployment and a residual for NILF.
   - White age categories: 16, 17, 18, 19, 20-24, 25-26, 27-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-62, 63-64, 65-67, 68-69, 70-74, 75+
   - Black age categories: 16-17, 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+
   - Other age categories: 16-44, 45+

   Iterate steps 1-3 six times.

The composite weighting steps are iterated six times. The iteration is needed since the execution of each step slightly imbalances the previous steps. After the race step, for example, weighted CPS estimates of the Hispanic labor force no longer match the controls used in the ethnicity step. After cycling through the process six times, CPS labor force estimates nearly match all three sets of controls. That is, the iterative raking process converges to the three sets of labor force controls.

Cell Collapsing -- Cells with few respondents have the potential of having large weight adjustments. The current second-stage weighting procedure has a collapsing algorithm for the ethnicity and race steps. For each cell in the ethnicity and race steps, a preliminary adjustment factor is computed. If the factor lies outside of the acceptable range 0.6-2.0 (or if the cell has no respondents), then the cell is collapsed with one or more adjacent age cells. The current composite weighting procedure also has a collapsing algorithm. For each cell in the ethnicity and race steps, preliminary adjustment factors are computed. Cells are collapsed if a labor force category has fewer than 10 responses or the factor falls outside of the acceptable range 0.7-1.3. This affects unemployment more than employment or not-in-labor force, but if a cell needs collapsing for one labor force category then it is collapsed for all. In a typical month about 20 cells are collapsed, mainly Black cells. The collapsing is an automated procedure and results can vary from month to month, but some Black cells are relatively small for unemployed and are collapsed in most months. [Ex: Assume 70% of the Civilian Noninstitutional Population (CNP) is in the Civilian Labor Force (CLF) and a 6% unemployment rate. A cell with 200 responses will have about 140 in the CLF: 132 employed, and 8 unemployed.]

Redesigned Composite Weighting

The redesigned composite weighting procedure to be implemented in January 2003 is also preceded by second-stage weighting procedure. California and New York are split into substate areas (Los Angeles-Long Beach metropolitan area and balance of California; New York City and balance of New York).

The redesigned second-stage weighting includes two new preliminary noniterated steps (a national-level coverage step and a state-level coverage step) that are followed by an iterative raking process that is similar to the current procedure. Specifications for the iterated steps follow. There will be ten iterations, although convergence to the unit usually only takes six to eight iterations. Rotation groups are paired for processing: (1,5), (2,6), (3,7), (4,8).

1. The first step of the iterative procedure, the state step, categorizes the observations into 318 cells: six gender x age cells for Los Angeles-Long Beach metropolitan area, the balance of California, New York City, the balance of New York, each of the remaining 48 states and the District of Columbia. Each cell is controlled to its CNP (each panel pair to CNP/4).
   - Age categories: 0-15, 16-44, and 45+

2. Following the state step, the ethnicity step categorizes the same observations into 52 cells: 26 Hispanic gender x age cells and 26 non-Hispanic gender x age cells. The Hispanic and non-Hispanic cells have identical age breaks as the Hispanic White and non-Hispanic Black cells in the national-level coverage step. Each cell is controlled to its CNP (each panel pair to CNP/4).
   - Hispanic and non-Hispanic age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, and 65+

3. The last step of the iterative procedure, the race step, categorizes the observations into 86 cells: 26 Black gender x age cells, 34 White gender x age cells, and 26 Asian and Residual combined gender x age cells. The Black, and Asian and Residual combined cells have identical age breaks as the ethnicity cells. The White cells have identical age breaks as the non-Hispanic White cells of the

national-level coverage step. Each cell is controlled to its CNP (each panel pair to CNP/4).

- Black, and Asian and Residual combined age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, and 65+
- White age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-62, 63-64, 65-69, 70-74, and 75+

The proposed procedure pairs rotation groups for second-stage weighting, whereas the current procedure processes each of the eight rotation groups separately. The pairings by month in sample or MIS are: incoming panels in MIS 1 and 5, panels in MIS 2 and 6, panels in MIS 3 and 7, and outgoing panels in MIS 4 and 8. Pairing panels for the redesigned second-stage weighting allows more demographic cell detail, but it is not possible to lump all panels together since that would ruin composite weighting. There are known month-in-sample biases associated with each pair that affect the composite estimation formula (the β_t term). Also, the stability of the Δ_t term would be adversely affected by combining all rotation groups together.

Composite estimation still applies only to persons 16 years of age and older. Using second-stage weights, composite estimates are made for employment and unemployment. The formulas for making composite estimates of employment and unemployment are unchanged and NILF is still derived as a residual. The composite estimates are then used as controls in the redesigned composite weighting procedure. For example, we anticipate about 880 responses per rotation group pair for the cell with White Males aged 30-34 – or about 3,520 responses for all rotation groups combined. Using second-stage weights, composite estimates are computed for White Male 30-34 employed and White Male 30-34 unemployed. White Male 30-34 NILF is obtained by subtracting these from the White Male 30-34 CNP control that was used for second-stage weighting in the race step. The two computed composite estimates and the NILF residual are then used as controls for composite weighting of the White Male aged 30-34 cell.

All eight rotation groups are combined for composite weighting. The iterative procedure is similar to the one for second-stage weighting. Step 1 uses the same 53 states/areas used in second-stage weighting. (Los Angeles-Long Beach metropolitan area; balance of California; New York City; balance of New York; the other 48 states; the District of Columbia.)

1. State Step -- Observations are categorized into 53 cells: a single cell for the District of Columbia, the Los Angeles-Long Beach metropolitan area, balance of California, new York City, balance of New York, and the remaining 48 states. Each cell is controlled to composite estimates of employment and unemployment and a residual for NILF.

2. Ethnicity Step -- Observations are categorized into 20 gender x age cells: 10 Hispanic cells and 10 non-Hispanic cells. Each cell is controlled to composite estimates of employment and unemployment and a residual for NILF.
   - Hispanic and non-Hispanic age categories: 16-19, 20-24, 25-34, 35-44, and 45+

3. Race Step -- Observations are categorized into 46 gender x age cells: 22 White cells, 14 Black cells, and 10 Asian and Residual cells. Each cell is controlled to composite estimates of employment and unemployment and a residual for NILF.
   - White age categories: 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+
   - Black age categories: 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45+
   - Asian and Residual age categories: 16-19, 20-24, 25-34, 35-44, 45+

Composite weighting cannot support the cell detail that is proposed for second-stage weighting. At first glance, combining all eight rotation groups together gives the appearance of larger cell sizes. However, sample counts of unemployed will turn out to be quite small if cells are defined too narrowly. This was anticipated and age ranges are "pre-collapsed" to avoid small unemployment counts.

Note that composite weighting partially unravels second-stage weighting. Take Black Male 35-44 as an example, where second-stage weights are controlled separately for ages 35-39 and 40-44. The two age ranges are collapsed for composite weighting. The composite weights will reproduce the Civilian Noninstitutional Population for Black Male 35-44, but not the CNP of the separate age groups. There are iterative procedures that could preserve all CNPs used for second-stage controls, but these would be more complex than the ones proposed here.

Changes in Composite Weighting

New Race --The descriptions above are given in terms of the new ethnicity/race data collection that allows reporting more than one race. The White, Black, and Asian race categories exclude multi-race reporters. The present Asian and Pacific Islander race category is split

into 2 categories: Asian, and Hawaiian and Other Pacific Islander. The Residual race category is comprised of: Native Hawaiian and Other Pacific Islander; American Indian, Aleut, and Eskimo; and Multi-Race. [To develop and test the new weighting procedures, data files with the present race categories were used: White, Black, Asian and Pacific Islander, and a residual that is American Indian, Aleut, and Eskimo. Note that this residual based on present race is quite a bit smaller than the new residual will be since it does not have Multi-Race or Native Hawaiian and Other Pacific Islander.]

No Cell Collapsing -- Software development and maintenance are simplified by eliminating an entire complex process. Cells are "pre-collapsed". A large majority of cells have more than 10 unemployed persons per panel pair. A cell monitoring system is needed to detect when survey conditions change and cells grow too small.

State Step Substate Areas -- California and New York have substate areas that are often included in tables with state data. Separate controls are introduced in the state steps for these substate areas: Los Angeles-Long Beach metropolitan area, balance of California, New York City, and balance of New York.

Ethnicity Step Demographic Detail -- Age groupings for Hispanics are shifted to be consistent with other age groupings. The proportion of Hispanics in the population is similar to the proportion of Blacks, but can only support 10 cells, gender by 5 age groupings. The same cell detail is used for Non-Hispanic, compared to the single cell of the present procedure.

Race Step 3 Demographic Detail -- Age detail is reduced in the White and Black groups. Asian is combined with Residual as a new race group in the race step and the 10-cell gender x age detail is the same as for Hispanic in the ethnicity step. Black can support 14 cells (7 age groups) and White can support 22 cells (11 age groups). The reduction in White and Black age detail was just a pre-collapsing to avoid low unemployment counts.

Consistently Defined Age Breaks -- White in the race step has 11 defined age breaks and these are logical collapsings of the 17 age breaks for White in the redesigned second-stage weighting. All other defined age breaks in the new composite weighting are logical collapsings of these in the sense that an age cell is never defined that splits one of the 11 age breaks. Convergence is assured to 10 national gender by age sets of labor force totals, all races and ethnicities combined. The age breaks used for the present weighting procedure have some inconsistencies and convergence is guaranteed to only 4 national gender by age sets of labor force totals, all races and ethnicities combined.

Consistency with Second-Stage Weighting – Changes for new composite weighting are consistent with changes made for new second-stage weighting. Major race categories are the same, and the age breaks used are consistent.

Ten Iterations -- Testing demonstrated that ten iterations are sufficient to get nearly exact convergence to control totals for all cells. The present weighting procedure has only 6 iterations and does not match controls as well as the new procedure will.

Unraveling Second-Stage Controls

Composite weighting partially unravels second-stage weighting. Most of the second-stage Civilian Noninstitutional Population controls are not matched when composite weights on CPS data files are summed.

- Composite weights do sum to second-stage CNP controls when the cells in the two procedures use the same age groups -- as is the case for some cells in the ethnicity and race steps.
- It is more common for second-stage age groups to be combined for composite weighting. When composite weights are summed for the second-stage cells, most CNP controls are missed by 1000 or more.
- Some second-stage CNP controls are missed by tens of thousands (maximum 33,648). Percentage-wise, occasional differences of up to 5% were found.

Fifteen months of CPS data (April 2000 – July 2001) were used to look for consistent biases over time between the sum of composite weights and the second-stage CNP controls. Two examples are given where two or more second-stage cells are combined for composite weighting. For the second-stage cells average differences were computed between composite weights and second-stage weights. If the mean difference equals the mean absolute difference for a cell, then the differences for all 15 months are in the same direction.

- Sometimes the plusses and minuses mostly cancelled out, as for non-Hispanic female age groups 25-29 and 30-34 that are combined for composite weighting.

	Approximate CNP	Mean Difference	Mean Absolute Difference
25-29	7,910,000	98	1883
30-34	8,685,000	-98	1883

- Sometimes there is a clear indication of systematic bias, as for some of the Hispanic male age groups 45-49, 50-54, 55-64, and 65+ that are combined for composite weighting.

	Approximate CNP	Mean Difference	Mean Absolute Difference
45-49	875,000	3176	3990
50-54	660,000	2030	2592
55-64	805,000	153	592
65+	715,000	-5360	6817

Any discrepancy at all from the "known" second-stage CNP controls can be of concern, but the indications of consistent bias are more troublesome. We plan to work on developing a more complex iterative weighting system that preserves both composite controls and second-stage controls.

For some analyses it will be better to use second-stage weights in preference to composite weights. This is particularly true for employment estimates within second-stage cells -- month-to-month changes will be more stable using second-stage weights for the cells than for composite weights that are developed after combining the cells.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

References:

Bailar, B.A. (1975). "The Effects of Rotation Group Bias on Estimates from Panel Surveys," **Journal of the American Statistical Association, 70**, 23-30.

Bureau of Labor Statistics and U.S. Census Bureau (2000), **Current Population Survey: Design and Methodology, Technical paper 63**. (www.bls.census.gov/cps/tp/tp63.htm)

Lent, J., S. Miller, P. Cantwell, and M. Duff (1999). "Effect of Composite Weights on Some Estimates from the Current Population Survey," **Journal of Official Statistics**.

Robison, Edwin (2001). "Proposal for Redesigned CPS Weighting, Second-Stage and Composite Procedures. " Prepared for the joint BLS/Census CPS Steering Committee.