# REDESIGNED CURRENT POPULATION SURVEY WEIGHTING

**Edwin L. Robison and Martha Duff, Bureau of Labor Statistics**
**Edwin 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**
Redesigned weighting for the monthly Current Population Survey (CPS) was implemented January 2003. New national and state coverage steps were added. A reconfigured Second-Stage iterative proportional fitting process forces weighted data to agree with independently derived population controls. A modified Composite Weighting procedure is another iterative process that forces weighted data to match sets of labor force controls generated with specialized composite formulas. The effect of each step and interactions among steps is analyzed. Particular emphasis is placed on the new state coverage step and the reconfigured Second-Stage state step. The steps afford much more detailed population control for states than prior procedures and dramatically stabilize monthly estimates for demographic subgroups within states.

**Background**
The Current Population Survey is jointly sponsored by the Bureau of Labor Statistics and the Bureau of the Census. It has a rotating panel survey design, and responses are obtained from about 50,000 households each month. The primary product of the CPS is labor force data for the Civilian Noninstitutional Population (CNP). 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 eight months and replaced by a returning panel. In adjacent months, six panels are in common. In a given month one panel each is being interviewed for the 1$^{st}$, 2$^{nd}$, 3$^{rd}$, 4$^{th}$, 5$^{th}$, 6$^{th}$, 7$^{th}$, and 8$^{th}$ time ("month-in-sample"). There are known biases in labor force data associated with the month-in-sample.

The BLS/Census CPS Weighting Group was formed in October 1999 to address weighting issues. A program of research was planned and executed that resulted in the implementation of new weighting procedures in January 2003. Of particular concern were revisions to the weighting needed because of changes in the race and ethnicity questions that were implemented for the first time in the January 2003 Current Population Survey. 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.

CPS weighting includes initial modules that account for the probability sampling of housing units, and adjust for noninterviewed households. Housing units are not selected from every locality; rather a first stage of sampling selects rural Primary Sampling Units to represent all rural PSUs. These are called non-self-representing or NSR PSUs; cities and other areas are called self-representing or SR PSUs. The selected NSR PSUs imperfectly represent all rural areas and First-Stage Ratio Adjustment computes Black versus non-Black ratio adjustments by state that are applied to the weights as a partial correction. Major changes were made to two iterative proportional fitting ("raking") procedures: Second-Stage Weighting and Composite Weighting.

Improvement efforts 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. Second-Stage Weighting uses monthly control totals of the Civilian Noninstitutional Population that are treated as fixed constants. The CNP controls are derived from models that update decennial demographic census totals using administrative data sources. Prior to 2003, successively computed ratio adjustments were made to weights in an iterative raking procedure so that estimates were forced to match three sets of CNP controls: 1) state controls; 2) national ethnicity x gender x age controls; and 3) national race x gender x age controls. The three steps were substantially modified based on a reevaluation of data analysis needs, convergence properties, and survey coverage. Fixed cells replace an on-the-fly collapsing technique, six gender x age controls are used for each state in the first step of the iterative procedure, Asian is added as a new race, and age categories are harmonized between successive steps of the iterative procedure to improve convergence. Two non-iterated coverage adjustment steps, national and state, were added. The new steps better account for known national interactions between ethnicity and race coverage and known differences in race coverage among the states.

Modifying the Composite Weighting procedure, immediately following Second-Stage Weighting in the weighting process, was also a high priority. For highly correlated items, an improved current-month estimate can be made by using previous months' data suitably adjusted with an estimate of change. A specialized composite estimation formula of this type has been used by the CPS for decades, and an innovative Composite Weighting procedure was implemented in 1998. The original and modified procedures use an iterative raking process to force weights to sum to three sets of controls. The controls are treated as constants in the process, but are computed exclusively from CPS sample data. Using composite formulas, prior CPS composite estimates (updated for change) are combined with current CPS estimates that have gone through Second-Stage Weighting. The controls are for employment, unemployment, and not-in-labor-force broken out into: 1) state controls; 2) national ethnicity x gender x age controls; and 3) national race x gender x age controls. The three steps were modified based on a reevaluation of convergence properties and interaction with Second-Stage Weighting.

In determining what changes to make to Second-Stage Weighting and 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 between the two procedures.
- Making control cell definitions more consistent across the state, ethnicity, and race steps, and also the noniterated Second-Stage coverage steps
- Pre-collapsing small cells to eliminate the need of the current "on-the-fly" collapsing algorithm that produces inconsistent results over time
- Providing more stable monthly estimates for population subgroups of interest to users (In particular, there was a request for demographic population controls within each state.)
- Confidence in the population controls for various age, race, ethnic, and geographic categories
- Possible changes in race reporting patterns over time
- Simplifying development and maintenance of the weighting software

Now that several months of "new" CPS data are available, we are continuing our weighting research. Of particular interest:

- Measure the improvement provided by the new Second-Stage coverage steps.
- Later Second-Stage steps partially "undo" the coverage steps. Test methods to somewhat reduce this.
- Measure the extent to which Composite Weighting "undoes" Second-Stage controls. Test methods to reduce this.
- Measure convergence. Prior to 2003 there were 6 iterations of the Second-Stage and composite steps; now there are 10.

This paper includes results we have obtained so far.

**Redesigned Second-Stage Weighting Procedure**

Second-Stage Weighting in the redesigned CPS weighting scheme includes two new preliminary steps, a national-level coverage step and a state-level coverage step, that are followed by an iterative raking procedure. The national-level coverage step was designed to account for the interaction between ethnicity and race, and the state-level coverage step was designed to account for differences in state race coverage relative to national coverage.

Several changes from pre-2003 procedures were introduced. In the state-level coverage step and the state step of the iterative procedure California and New York are split into substate areas: Los Angeles-Long Beach metropolitan area and the balance of California, and New York City and the balance of New York, respectively. Age and Black/non-Black detail is added to the state coverage step. The iterated state step includes gender and age breaks. In the national-level coverage step and the race step of the iteration procedure, Asian is introduced as a new race. All iterated steps now have a major age break separating 0-15 year olds from the 16+ population (important since Composite Weighting applies only to CNP16+). Each adjustment of the redesigned procedure consists of a fixed number of cells that all have adequate sample sizes -- a collapsing algorithm was eliminated. Each step, excluding specific cells of the state coverage step, has population control totals that are estimates of CNP/4, one-fourth of the Civilian Noninstitutional Population. Formerly each panel was treated separately, but now the eight monthly panels are paired to increase cell counts, allowing more demographic detail.

The new race/ethnicity data collection implemented January 2003 allows multi-race reporting. Also, it is now possible to distinguish Asians from Native Hawaiians and Other Pacific Islanders. White, Black, and Asian cells in the cell specifications that follow exclude multi-race reporters. The Residual race includes observations not categorized strictly as Asian, Black, or White: Native Hawaiian and Other Pacific Islander; American Indian, Aleut, and Eskimo; and Multi-Race.

Cell definitions are consistent with the needs of the Composite Weighting procedure that comes after Second-Stage Weighting. In particular, age breaks are defined consistently between the two procedures in order to minimize the extent to which Composite Weighting "undoes" the Second-Stage population controlling.

Specifications follow for cells of the national-level coverage step (A), the state-level coverage step (B), and the three steps (1-3) of the redesigned iterative procedure.

A. The non-iterated national-level coverage step, adjusting for subpopulations prone to under/over coverage, categorizes the CPS sample observations into 126 cells: 26 Hispanic White gender x age cells, four Hispanic non-White gender x age cells, 18 non-Hispanic Asian gender x age cells, 26 non-Hispanic Black gender x age cells, 34 non-Hispanic White gender x age cells, and 18 non-Hispanic Residual gender x age cells. The Hispanic White and non-Hispanic Black cells have identical age breaks, and the non-Hispanic Asian and non-Hispanic Residual cells have identical age breaks.
   - Hispanic White and non-Hispanic Black 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+
   - Hispanic non-White age categories: 0-15, and 16+
   - Non-Hispanic Asian and non-Hispanic Residual age categories: 0-4, 5-9, 10-15, 16-24, 25-34, 35-44, 45-54, 55-64, and 65+
   - Non-Hispanic 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+

B. The non-iterated state-level coverage step allows more detailed controlling for states/substates with larger numbers of persons of Black race in the sample. All 515 defined cells are controlled to CNP. In some cells the panels are paired and controlled to CNP/4, but for others the controlling is only to CNP with all eight panels combined.

   - Non-Black – In all states/substates six gender x age cells (0-15, 16-44, 45+) are defined. Each of the 318 cells is controlled to its CNP. Except for the District of Columbia, each panel is controlled to its CNP/4.
   - Black – In 26 states/substates six gender x age cells (0-15, 16-44, 45+) are defined. The eight panels are combined and each of the 156 cells is controlled to its CNP: AL, AR, CT, DE, FL, GA, IL, LA, MD, MA, MI, MO, MS, NJ, NC, OH, PA, SC, TN, TX, VA and the District of Columbia, Los Angeles-Long Beach metropolitan area, the balance of California, New York City, the balance of New York.
   - Black – In 14 of the states with smaller Black race populations two gender cells are defined, age combined. Panels are combined and each of the 28 cells is controlled to its CNP: AK, AZ, CO, KY, OK, IN, KS, MN, NE, NV, RI, WA, WV, and WI.
   - Black – For the remaining 13 states, those with the smallest Black race population, one cell is defined, gender and age combined. Panels are combined and each of the 13 cells is controlled to its CNP: HI, IA, ID, ME, MT, NH, NM, ND, OR, SD, UT, VT, and WY.

1. State Step (iterated) -- Observations are categorized 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. Ethnicity Step (iterated) -- Observations are categorized 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. Race Step (iterated) – Observations are categorized 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 steps are iterated – ten times – separately for each of the four CPS panel pairings. For each cell k of each step, a simple adjustment is computed for each of the panel pairs 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.

$$ADJ_{ki} = (CNPk/4)/ \sum_j w_{kij}$$
$$w_{kij}' = ADJ_{ki} * w_{kij}$$

Several iterations of steps 1-3 are needed, since each step slightly imbalances the previous steps. After the race step, for example, weighted CPS estimates of population for a given rotation group pair no longer match the population control by ethnicity/gender/age or by state. After cycling through the process ten times, CPS estimates of population for each rotation group pair nearly match all three sets of controls. That is, the iterative raking process converges to the three sets of population controls. (Also, if you use all eight panels, CPS estimates of population nearly match the desired Civilian Noninstitutional Populations.)

**Findings on Second-Stage Convergence**

Convergence to the unit usually takes only six to eight iterations. We verified that changing the order of the steps in the iterative procedure has no effect on the final results since the set of equations tend to have a unique solution.

Convergence -- Consistently defined age breaks are, more than any other factor, the key to fast convergence in the iterative procedure. For the redesign, White in the race step has 17 age breaks, and all other defined age breaks are logical collapsings of those 17. The "magic number" is the 13 age breaks used in the ethnicity step and in the race step for Black and Residual. We would have liked to further split out Asian in the race step -- but neither Asian nor the very small residual could support 13 age breaks -- and in testing convergence slowed to a crawl. Incidentally, convergence is now assured to 26 national gender-by-age population totals. Inconsistencies slowed convergence for the pre-2003 procedure, and in six iterations several population controls were missed by hundreds. Due to inconsistencies, only 10 national gender-by-age population controls were matched. The new coverage steps also affect convergence – the national coverage step speeds convergence, but the state coverage step slightly slows convergence.

**New Coverage Steps**

National-Level Coverage Step – This non-iterated step helps correct for interactions between race and ethnicity coverage that proved impossible to address in our iterative procedure. For example, research discovered gross undercoverage of Non-Black Hispanics that can be corrected for in this step but not in the iterative steps. Without the national-level coverage step non-Hispanic Asians (shown in Figure 1 below), non-Hispanic Blacks, non-Hispanic Residuals, and Hispanic Whites tend to be overestimated at the end of the Second-Stage iterative procedure; whereas, non-Hispanic Whites, and Hispanic Asians, Hispanic Blacks, and Hispanic Residuals tend to be underestimated.

Figure 1 visually illustrates the improvement for non-Hispanic Asians using January-June 2003 data. Although Asian controls were unworkable in the iterative race step, a reasonable degree of control is made possible by the coverage step. The step has 18 controls for non-Hispanic Asians (gender by 9 age breaks; over 95 percent of all Asians included). Later steps partially "undo" the control and the "With A" box plot summarize the relative difference between the final weighted estimates for non-Hispanic Asians and the 18 controls for each month. These are much closer to the ideal (0 -- when the estimate equals the control) than the "Without A" box plot that summarizes the relative differences from the controls when the step is omitted. The same general picture emerges every month for all ethnicity/race subpopulations that are specifically used in the national-level coverage step.

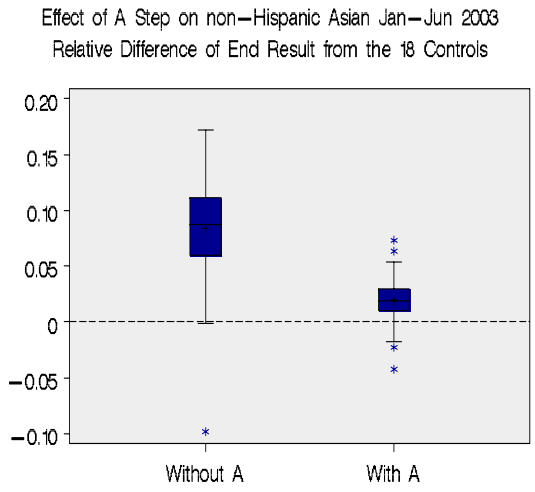


**Figure 1.**

State-Level Coverage Step – This non-iterated step adjusts for state differences in gender/age/race coverage. It proved impossible to include race in an iterated state step. Figure 2 visually illustrates the value of the state-level coverage step using January-June 2003 data. The "With B" box plot compares final weighted estimates to controls for 197 Black cells. The "Without B" box plot shows the same comparison when the state coverage step is omitted. Without the state-level coverage step, some estimates are quite far off from the controls. With the step, almost all estimates are within 5 percent of the controls at the end of the Second-Stage Weighting procedure. (In further testing, it has been found that repeating the state coverage step a second time shrinks differences from controls by about half without appreciably slowing convergence.)

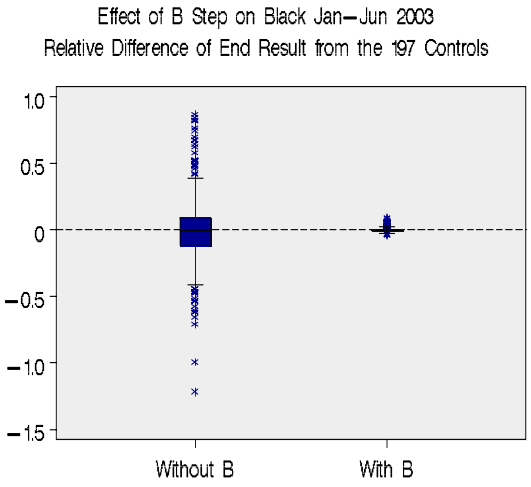


**Figure 2.**

**Redesigned Composite Weighting Procedure**

In general outline, the redesigned Composite Weighting is much like the pre-2003 methodology. The procedure applies only to persons 16 years of age or older (CNP16+) and all panels are combined when weight adjustments are applied within cells. Using Second-Stage weights, composite estimates are made of employment and unemployment for demographic groups, and not-in-labor-force (NILF) is derived as a residual. (For a given demographic group, the three sum up to a Civilian Noninstitutional Population figure that is treated as a known constant.) The composite estimates are then used as controls for the Composite Weighting procedure. The formulas for making composite estimates of employment and unemployment are unchanged. They 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 $\Delta_t$ developed from the six continuing panels between last month and this month (specified by month-in-sample 2,3,4,6,7,8 for month t in the formulas). Usually $\beta_t$ 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.

- 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
  $Y^c_t = .3Y^{ss}_t + .7(Y^c_{t-1} + \Delta_t) + .4\beta_t$
- $\Delta_t = (4/3)\Sigma( x_{t,i} - x_{t-1,i-1})$ sum over i=2,3,4,6,7,8
- $\beta_t = x_{t,1} + x_{t,5} - (1/3)\Sigma( x_{t,i})$ sum over i=2,3,4,6,7,8

All panels are combined for Composite Weighting. The 3-step procedure with 10 iterations is similar to the pre-2003 procedure. It has the same 3-step structure as Second-Stage Weighting, and step 1 uses the same 53 states/areas. Important changes were made to harmonize the age categories between steps and with the Second-Stage.

1. State Step (iterated) -- Observations are categorized into 53 cells: a single cell for the 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 composite estimates of employment and unemployment and a residual NILF.

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

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

Composite Weighting partially "undoes," unravels, or deconstructs Second-Stage Weighting. That is mainly because it cannot support the same demographic cell detail as Second-Stage Weighting. Combining all eight panels together helps create larger cell sizes for Composite Weighting. However, sample counts of unemployed get quite small for cells that are defined too narrowly. This was anticipated, the counts carefully reviewed in developmental research, and a collapsing algorithm for small cells was eliminated. The box plots in Figure 3 summarize for January-June 2003 the relative differences between final weighted estimates of the controls (made after Composite Weighting) and the actual Second-Stage controls. No control in any month is unraveled by more than 2%.

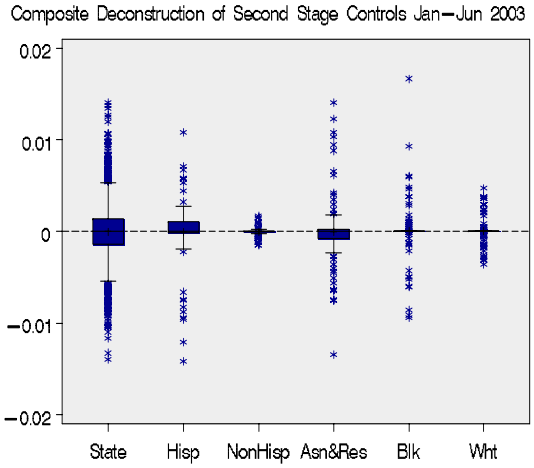


**Figure 3.**

**Reducing Composite Deconstruction of Second-Stage Controls**

Promising additional research is continuing on modifying the iterative weighting procedures in order to minimize the extent to which Composite Weighting unravels the Second-Stage controls. Ideally, one set of weights would be produced that simultaneously matched both the Second-Stage CNP controls and the composite controls.

Figure 4 shows how much can be achieved by simply repeating the weighting procedure a 2nd, 3rd, and 4th time. Each repetition includes Second-Stage Weighting (10 iterations without the initial coverage steps, using the same CNP controls) and Composite Weighting (10 iterations, using the same controls). The box plots summarize the January-June 2003 unraveling of Second-Stage state-step controls. The worst outliers dramatically shrink from almost 1.5% to about .2% when the weighting procedure is run a 2nd time. Outliers shrink even further to about .03% on the 3rd time and to half of that on the 4th time. The system appears to converge.

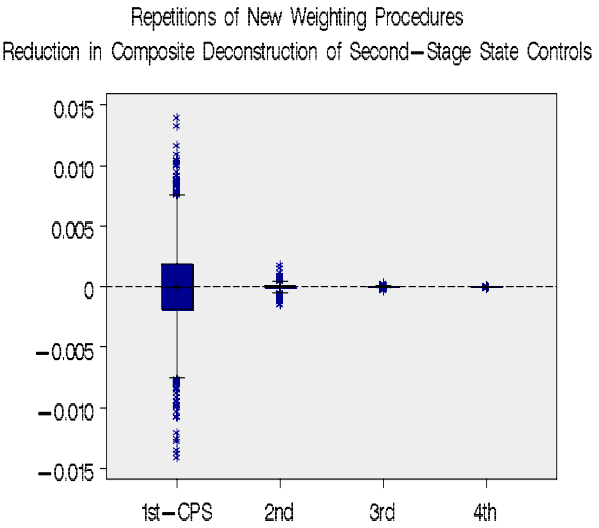


**Figure 4.**

**Disclaimer**

*Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics or the Bureau of the Census.*

References:

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**.

Bureau of Labor Statistics and U.S. Census Bureau (2000), **Current Population Survey: Design and Methodology, Technical Paper 63**. (www.census.gov/prod/2000pubs/tp63rv.pdf)

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

Robison, Edwin, Martha Duff, Brandon Schneider, and Harland Shoemaker (2002). "Redesign of Current Population Survey Raking to Control Totals." Presented at the 2002 ASA, Section on Survey Research Methods.

Robison, Edwin, Martha Duff, Brandon Schneider, and Harland Shoemaker (2002). "Redesign of Current Population Survey Composite Weighting." Presented at the 2002 ASA, Section on Survey Research Methods.