Weighting the overlap sample obtained from two Tobacco Use Supplements to the Current Population Survey

William W. Davis, Anne M. Hartman, James T. Gibson

October 2007

1. Introduction

The Current Population Survey (CPS) is a monthly labor force survey conducted by the Census Bureau (CB) and sponsored by the Bureau of Labor Statistics (BLS) in approximately 55,000 interviewed households across the country. Census Bureau staff collected the series of National Cancer Institute (NCI) sponsored Tobacco Use Supplements (TUS) to the CPS initiating in September 1992. Beginning in 2001-02, the Centers for Disease Control and Prevention (CDC) has co-sponsored these surveys with the NCI. The most recent series of Tobacco Use Supplements was conducted in May 2006, August 2006 and January 2007. Census Bureau staff also conducted a Tobacco Use Special Cessation Supplement (TUSCS) to the February, June and November 2003 CPS (US Department of Commerce, Census Bureau, 2006). This data is available for public use along with data from September 1992, January 1993 and May 1993 through June 2001, November 2001 and February 2002 (see http://riskfactor.cancer.gov/studies/tus-cps/).

A unique feature of the CPS is its panel design where each household in the sample is surveyed for four consecutive months and then for four more consecutive months nine months later (see, Current Population Survey, 2002). Due to this sampling strategy persons who were in their 1st, 2nd, or 3rd month in sample in February 2002 when the TUS data was collected were potentially also in the February 2003 sample for panel months #5, 6 and 7 when the TUSCS-CPS was fielded. The 4th month in sample in Feb. 2002/8th month in sample in Feb. 2003 weren't given the TUS (typically all 8 panels are given the TUS in *other* months of data collection). We refer to those who responded to both the February 2002 TUS and the February 2003 TUSCS as the overlap sample. The responses to the overlap sample can be analyzed as one-year longitudinal study with a representative sample of the U.S. and hence furnishes a unique opportunity for data analysis. In addition, if retrospective information asked in February 2002 about the period of time between February 2001 and February 2002 is included, then potentially a two-year period can be examined.

Each person in the overlap sample had the following statistical weights for both February 2002 and February 2003:

- Full sample supplement non-response weights
- Replicate supplement non-response weights (80)
- Full sample supplement self-response weights
- Replicate supplement self-response weights (80)

The self-response weights are used for analysis of items involving self-response only (that is no proxy responses). The replicate weights incorporate the complex design of the TUS-CPS sample, and their use allows accurate estimation of standard errors. The CPS method to derive the full-sample weights is described in Current Population Survey (2002, Chapter 10). The CPS weighting method involves both ratio adjustment and raking using cells defined by geography (states), ethnicity (Hispanic vs. not), races (White, Black, and Other), gender, and age categories.

Either the February 2002 or the February 2003 statistical weights could be used to construct (weighted) population estimates. However, the overlap sample is a subset of both of these so that overlap sample analyses using either the February 2002 or the February 2003 weights may be biased. For this reason, we chose to re-weight the overlap sample to adjust for differential non-response and non-coverage by gender, race/ethnicity, age, and geography. We used the same general method to adjust the four sets of full sample and replicate weights shown above.

We could have applied the weight adjustment to either the February 2002 or the February 2003 weights. The statistical weights for the February 2002 sample were based on control totals obtained by projecting Census 1990 totals while the statistical weights for the February 2003 sample were based on control totals obtained from the more recent Census 2000. Because of this fact, we felt that the statistical weights for February 2003 were more accurate and adjusted these rather than the February 2002 weights.

In Section 2, we summarize the overlap sample characteristics and use these to justify the weight modification. In Section 3, we describe the weighting methodology and show the impact of the weight modification in Section 4

2. Overlap Sample Characteristics

Table 1 shows number of respondents and weighted totals (using the full sample supplement non-response weights) for the February 2003 TUSCS and the overlap sample by gender, Census region, age, and race/ethnicity (specified as Hispanic, non-Hispanic (NH) White, NH Black, and NH Other). We have restricted attention to survey respondents who were age 15 and older, the age range that is interviewed for the CPS and TUS (TUSCS). Table 1 shows that the total number of overlap respondents was 22,598 and that roughly one third (32.8%) of the 68,954 February 2003 respondents were in the overlap sample.

Table 1. Weighted and unweighted counts for the February 2003 TUSCS for the complete and overlap sample

	Unweighted Counts				Weighted Counts		
	Overlap	All	Overlap Percent		Overlap	All	Overlap Percent
All	22,598	68,954	32.8%		71,752,091	224,088,640	32.0%
Males	10,440	32,712	31.9%		33,589,980	107,871,180	31.1%
Females	12,158	36,242	33.5%		38,162,111	116,217,460	32.8%
Region							
North East	4,945	14,922	33.1%		14,046,554	43,147,939	32.6%
North Central	5,925	17,231	34.4%		17,418,247	50,780,180	34.3%
South	6,552	19,809	33.1%		25,562,187	79,738,280	32.1%
West	5,176	16,992	30.5%		14,725,103	50,422,242	29.2%
Age							
15-24	2,648	11,233	23.6%		9,287,606	39,715,297	23.4%
25-34	3,170	11,255	28.2%		10,643,890	38,935,646	27.3%
35-44	4,743	13,644	34.8%		14,890,812	43,839,449	34.0%
45-54	4,631	12,770	36.3%		14,574,243	40,214,366	36.2%
55-64	3,434	8,803	39.0%		10,394,745	27,251,655	38.1%
65+	3,972	11,249	35.3%		11,960,795	34,132,226	35.0%
Race/ethnicity							
Hispanic	1,771	6,684	26.5%		7,309,211	27,812,152	26.3%
NH White	17,947	52,152	34.4%		53,784,871	157,866,726	34.1%
NH Black	1,844	6,129	30.1%		7,442,553	25,454,962	29.2%
NH Other	1,036	3,989	26.0%		3,215,456	12,954,800	24.8%

The TUS-CPS recommendation is to group all respondents from three separate monthly surveys together to make state estimates. The number of overlap respondents is roughly 10 percent of the number in a typical three month combined full supplement sample so the sample is not sufficient to make estimates for all states (The state overlap samples range from 202 in the District of Columbia to 1,369 in California). Thus, we used the four Census regions (not the 51 states including the District of Columbia) as the geographical area of interest when we modified the statistical weights. Thus, we do not recommend making state estimates or summaries in the analysis of the overlap sample.

Table 1 shows that the ratio of the weighted overlap counts to the total counts (using the full-sample non-response weights) was 32.0%. Table 1 also shows the ratio for gender, region, age, and race/ethnicity groups. If the overlap percentage for the weighted counts for a row of Table 1 differs significantly from 32.0%, it suggests differential non-response and/or non-coverage. The smallest ratios are 29.2% for the West region, 23.4% for the youngest age category, and 24.8% for NH Other. If no weighting adjustment is made, these groups will be under-represented in a weighted analysis using the entire overlap sample. Similarly, those groups with a ratio higher than 32.0% (such as the older age-groups) will be over-represented in a weighted analysis using the entire overlap sample. Thus, we modified the overlap sample weights to correct as much as possible for differential non-response and non-coverage in the overlap sample.

Table 2 provides a similar summary to Table 1 but using only the self-respondents and the full sample supplement self-response weights. It supports the need to modify the self-response weights to correct for differential non-response and non-coverage in the overlap sample. Table 2 shows that of the 54,306 self-respondents in the February 2003 TUSCS the total number in the overlap sample was 15,846 (29.2%). Table 2 shows that the ratio of the weighted overlap counts to the weighted total counts (using the full-sample supplement self-response weights) was 27.5%. If no weighting adjustment were made, groups would be over (under) represented if their ratio was larger (smaller) than 27.5%. The smallest ratios for the weighted self-response counts were 25.1% for the West region, 16.1% for the youngest age category, and 20.4% for NH Other. The three highest age groups had the largest ratios.

Table 2. Weighted and unweighted counts for self-response for the February 2003 TUSCS for the complete and overlap sample

	Unweighted Self-response Counts				Weighted Self-response Counts		
	Overlap	All	Overlap Percent		Overlap	All	Overlap Percent
All	15,846	54,306	29.2%		61,644,786	224,088,640	27.5%
Males	6,516	24,099	27.0%		27,600,780	107,789,462	25.6%
Females	9,330	30,207	30.9%		34,044,007	116,299,178	29.3%
Region							
North East	3,290	11,180	29.4%		11,940,449	43,045,520	27.7%
North Central	4,303	13,948	30.9%		15,117,466	50,768,761	29.8%
South	4,571	15,696	29.1%		21,911,371	79,841,805	27.4%
West	3,682	13,482	27.3%		12,675,500	50,432,554	25.1%
Age							
15-24	1,143	7,016	16.3%		6,383,301	39,715,297	16.1%
25-34	2,225	9,092	24.5%		9,103,879	38,935,646	23.4%
35-44	3,385	10,918	31.0%		12,957,745	43,839,449	29.6%
45-54	3,282	10,374	31.6%		12,633,708	40,214,366	31.4%
55-64	2,557	7,288	35.1%		9,243,595	27,251,656	33.9%
65+	3,254	9,618	33.8%		11,322,557	34,132,226	33.2%
Race/ethnicity							
Hispanic	1,091	4,923	22.2%		5,966,826	27,811,564	21.5%
NH White	12,849	41,660	30.8%		46,689,945	157,868,841	29.6%
NH Black	1,262	4,756	26.5%		6,336,500	25,434,481	24.9%
NH Other	644	2,967	21.7%		2,651,516	12,973,755	20.4%

3. Weighting methodology

We modify the February 2003 weights using ratio adjustment based on geography, race/ethnicity, gender, and age categories using the following notation:

- Census region g=1,..,4 (North East, North Central, South, and West)
- Sex s=1, 2 (male, female)
- Race/ethnicity r=1,..,4 (Hispanic, non-Hispanic (NH) White, NH Black, NH Other)
- Age category a=1,…,A

For each region, gender race/ethnicity, and age we define mutually exclusive and exhaustive groups, B_{gsra}.. The number of age groups chosen is labeled as A_{gsr} so that the total number of groups formed is $A = \sum_{g,s,r} A_{gsr}$. Within each of these groups we calculate the ratio of the sum of the weights for the February 2003 sample to the sum of the weights in the overlap sample. This ratio is used to modify the statistical weight of all the overlap respondents in this group. If O denotes the set of overlap sample respondents (n=22,598), then for the i^{th} individual the adjusted full-sample weight, $w_i^{'}$, is the product of the full-sample weight, w_i , and the ratio adjustment. In symbols we have

$$w_i^{'} = w_i r_{gsra} \qquad i \in B_{gsra} \cap O \tag{1}$$

where the ratio-adjustment is

$$r_{gsra} = \sum_{i \in B_{gsra}} w_i \Big/ \sum_{i \in B_{gsra} \cap O} w_i \tag{2}$$

This technique insures that the sum of weights of the overlap sample matches the sum of the weights for the entire sample in each of the *A* groups as follows

$$\sum_{i \in B_{gsra} \cap O} w_i^{'} = \sum_{i \in B_{gsra}} w_i \tag{3}$$

We chose the number of age categories, A_{gsr}, so that there would be a sufficient overlap sample to estimate the ratio in (2). Table 3 shows the criterion used to choose the number of age groups as a function of the overlap sample size and the age range for the age cells are shown in Table 4. The age ranges were chose to match age-range of the cells of the CPS statistical weighting procedure as much as possible. For example, for a overlap sample size of 110, table 3 shows that 2 age groups were used, and table 4 shows that the age range of these groups was 15-44 and 45+.

Table 3. Number of groups as a function of the overlap sample size

Overlap sample size	Number of age groups
N<120	2
120≤N<240	4
240≤N<360	6
360≤N<720	10
N>720	14

Table 4. Definition of age ranges as a function of the number of age groups

	Total number of age groups				
Age range	14	10	6	4	2
15-19	1	1	1	1	1
20-24	2	2			
25-29	3	3	2		
30-34	4	4		2	
35-39	5	5	3		
40-44	6	6			
45-49	7	7	4	3	2
50-54	8	8			
55-59	9	9	5	4	
60-62	10				
63-64	11				
65-69	12	10	6		
70-74	13				
75+	14				

4. Weighting results

Table 5 shows the overlap sample size and the number of age groups as a function for the 32 groups defined by region, sex, and race/ethnicity ordered by overlap sample size for both the non-response and the self-response weights. For the non-response weights, the table shows that 10 of the groups have only 2 age-adjustment categories, 6 have four categories, 4 have 6 categories, 4 have 10 and 8 have 14 categories. The table shows that the total number of groups is A=220 (=10*2+6*4+…+8*14) for non-response and A=190 for self-response. The table shows that the all eight NH White groups had 14 age groups for both self and non-response and consequently had 14 age groups, where the number of groups was determined from table3.

Equation (3) implies equality of sums for A cells defined by four dimensions. The equality of sums for these cells implies equality also for lower dimensional marginal sums such as the the 32 groups of table 5 defined by region, sex, and race/ethnicity. For either self- or non-response, these sums can be obtained by summing over the age-categories defined in tables 4 and 5 using either the regular or modified weights.

Table 5. Number of age groups by region, race/ethnicity and gender ordered by the overlap non-response sample size

			Non-response		Self-response	
Census region	Race/ethnicity	Gender	Overlap sample size	Number of age groups	Overlap sample size	Number of age groups
Northeast	NH Other	Male	48	2	30	2
West	NH Black	Male	58	2	39	2
Northeast	NH Other	Female	60	2	34	2
North Central	Hispanic	Female	74	2	50	2
North Central	Hispanic	Male	78	2	43	2
West	NH Black	Female	79	2	65	2
North Central	NH Other	Male	80	2	44	2
South	NH Other	Male	83	2	53	2
North Central	NH Other	Female	90	2	68	2
South	NH Other	Female	113	2	83	2
Northeast	NH Black	Male	121	4	66	2
Northeast	Hispanic	Male	132	4	51	2
North Central	NH Black	Male	147	4	79	2
Northeast	Hispanic	Female	164	4	88	2
Northeast	NH Black	Female	198	4	145	4
North Central	NH Black	Female	218	4	166	4
West	NH Other	Male	263	6	140	4
South	Hispanic	Male	264	6	167	4
South	Hispanic	Female	275	6	204	4
West	NH Other	Female	299	6	192	4
West	Hispanic	Male	377	10	203	4
West	Hispanic	Female	407	10	285	6
South	NH Black	Male	419	10	247	6
South	NH Black	Female	604	10	455	10
West	NH White	Male	1729	14	1190	14
Northeast	NH White	Male	1963	14	1187	14
West	NH White	Female	1964	14	1568	14
South	NH White	Male	2207	14	1357	14
Northeast	NH White	Female	2259	14	1689	14
North Central	NH White	Male	2471	14	1620	14
South	NH White	Female	2587	14	2005	14
North Central	NH White	Female	2767	14	2233	14
Total			22,598	220	15,846	190

For each of the 32 groups in Table 4, Table 6 shows the following for both self- and non-response weights

- coefficient of variation (CV) of the full-sample weights
- coefficient of variation (CV) of the modified weights
- ratio of the CVs for the modified to the full-sample weights

The CV is a useful summary of the weights, because the length of a confidence interval for the (weighted) mean increases linearly with the CV of the weights (e.g., Korn and Graubard, 1999, Chapter 4). Thus, the modified weights correct for the potential bias, but yield increased length confidence intervals (for those categories where the ratio is more than 1). Table 6 shows that the ratio of CVs is larger than 1 for 29 of the 32 groups for non-response (range 0.90 to 1.40) and 28 of the 32 groups for self-response (range 0.82 to 1.42).

The CVs of the entire sample for the original and modified full sample non-response weights are 59.6% and 69.5% respectively with a ratio of 1.17. This shows a 17% increase in the length of confidence intervals using the entire population, where the increase is due to the statistical weight adjustment to eliminate the potential overlap sample bias. Similarly, the CVs of the entire sample for the original and modified full sample self-response weights are 65.7% and 86.7% respectively with a ratio of 1.32. This shows a 32% increase in the length of confidence intervals using the entire population of self-respondents.

The ratio adjustment procedure described in equations (1) and (2) was applied to the 80 self- and non-response replicate weights yielding 80 modified self- and non-response replicate weights. The summary statistics for the coefficient of variation of these replicate weights are shown in Table 7. In general, the CVs of the replicates are larger than the full-sample. For example, the original full sample non-response weight CV is 59.6% while the replicate Non-response CVs range between 67.7% and 84.6% with a mean of 82.9%. The table shows that the mean ratio for the non-response weights is 1.11 and the mean ratio for the self-response weights is 1.21. This shows that the confidence intervals obtained from the modified non-response weights would be approximately 11% longer than those obtained using the regular weights. These increased lengths seems to be a small price to pay in order to obtain a reduction in bias of the estimates.

Table 6. Coefficient of variation expressed as a percent of the original and the modified non-response and self-response statistical weights by gender, region and race/ethnicity

			Cofficient of variation for Non-response weights			Cofficient of variation for Self-response weight		
Region	Race/ethnicity	Sex	Original	Modified	ratio	Original	Modified	ratio
North Central	Hispanic	Female	37.3	42.1	1.13	44.8	58.8	1.31
North Central	NH Black	Female	35.0	46.6	1.33	38.0	69.4	1.83
North Central	NH Other	Female	72.2	66.1	0.92	78.4	75.3	0.96
North Central	NH White	Female	52.4	58.5	1.12	58.3	75.3	1.29
North East	Hispanic	Female	42.7	49.3	1.15	48.1	56.2	1.17
North East	NH Black	Female	33.6	46.2	1.38	40.7	63.8	1.57
North East	NH Other	Female	53.5	53.8	1.01	61.8	57.2	0.93
North East	NH White	Female	61.2	64.3	1.05	67.1	80.2	1.19
South	Hispanic	Female	30.0	40.4	1.35	31.8	39.9	1.25
South	NH Black	Female	53.8	65.1	1.21	56.2	77.3	1.38
South	NH Other	Female	51.6	66.4	1.29	59.0	81.3	1.38
South	NH White	Female	42.6	49.5	1.16	48.0	64.7	1.35
West	Hispanic	Female	56.4	59.5	1.06	57.4	63.9	1.11
West	NH Black	Female	55.7	60.4	1.08	57.5	68.7	1.19
West	NH Other	Female	94.6	94.4	1.00	94.7	106.6	1.13
West	NH White	Female	76.7	82.2	1.07	81.4	105.9	1.30
North Central	Hispanic	Male	39.8	53.5	1.34	43.1	61.3	1.42
North Central	NH Black	Male	41.9	46.5	1.11	50.0	50.9	1.02
North Central	NH Other	Male	72.8	88.9	1.22	73.9	82.2	1.11
North Central	NH White	Male	55.1	59.9	1.09	59.5	72.4	1.22
North East	Hispanic	Male	43.5	44.5	1.02	46.3	38.1	0.82
North East	NH Black	Male	34.0	42.3	1.24	47.6	60.6	1.27
North East	NH Other	Male	58.7	53.0	0.90	60.2	71.2	1.18
North East	NH White	Male	60.7	66.7	1.10	66.4	89.3	1.34
South	Hispanic	Male	34.8	48.8	1.40	41.1	49.7	1.21
South	NH Black	Male	55.8	64.2	1.15	66.5	80.1	1.21
South	NH Other	Male	52.6	57.3	1.09	56.3	58.9	1.05
South	NH White	Male	43.4	49.5	1.14	49.7	63.9	1.29
West	Hispanic	Male	60.4	67.9	1.13	62.7	74.8	1.19
West	NH Black	Male	68.8	64.1	0.93	78.6	75.1	0.96
West	NH Other	Male	95.8	108.1	1.13	94.2	123.1	1.31
West	NH White	Male	77.6	83.2	1.07	81.4	93.7	1.15

Table 7. Summary statistics for the coefficient of variation (CV) of the non-response and self-response replicate weights: original and modified

		CV Mean	CV Standard Deviation	CV Minimum	CV Maximum
Non-response replicate weights	Original	82.8%	3.0%	67.7%	84.6%
	Modified	91.7%	2.9%	75.4%	93.8%
	Ratio	1.11	0.01	1.08	1.15
Self-response replicate weights	Original	88.3%	3.0%	71.4%	89.9%
	Modified	108.0%	2.9%	92.3%	111.4%
	Ratio	1.22	0.02	1.19	1.29

References

Current Population Survey "Design and Methodology. Technical Paper 63RV, May 2002, U.S. Department of Labor, Bureau of Labor Statistics.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons.

US Department of Commerce, Census Bureau 2006, National Cancer Institute and Centers for Disease Control and Prevention co-sponsored Tobacco Use Special Cessation Supplement to the Current Population Survey 2003 – http://riskfactor.cancer.gov/studies/tus-cps/.