ATTACHMENT J

Source and Accuracy of the Data for the
May 2004 CPS Microdata File for Work Schedules and Work at Home

SOURCE OF DATA

The data for this microdata file come from the May 2004 Current Population Survey (CPS). The May survey uses two sets of questions, the basic CPS given every month and the May 2004 supplement. The CPS, sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics, is the country's primary source of labor force statistics for the entire population.

Basic CPS. The monthly CPS collects primarily labor force data about the civilian noninstitutional population living in the United States. Interviewers ask questions concerning labor force participation about each member 15 years old and over in sample households.

The CPS uses a multistage probability sample based on the results of the decennial census, with coverage in all 50 states and the District of Columbia. When files from the most recent decennial census become available, the Census Bureau gradually introduces a new sample design for the CPS[1].

In April 2004, the Census Bureau began phasing out the 1990 sample and replacing it with the 2000 sample, creating a mixed sampling frame. Two simultaneous changes will occur during this phase-in period. First, primary sampling units (PSUs)[2] selected for only the 2000 design will gradually replace those selected for the 1990 design. This will involve 10 percent of the sample. Second, within PSUs selected for both the 1990 and 2000 designs, sample households from the 2000 design will gradually replace sample households from the 1990 design. This will involve about 90 percent of the entire sample. By July 2005, the new sample design will be completely implemented, and the sample will come entirely from Census 2000 files.

In the first stage of the sampling process, PSUs are selected for sample. In the 1990 design, the United States was divided into 2,007 PSUs. These were then grouped into 754 strata, and one PSU was selected for sample from each stratum. In the 2000 sample design, the United States is divided into 2,025 PSUs. These PSUs are grouped into 824 strata. Within each stratum, a single PSU is chosen for the sample, with its probability of selection proportional to its population as of the most recent decennial census. This PSU represents the entire stratum from which it was selected. In the case of strata consisting of only one PSU, the PSU is chosen with certainty.

The 1990 design and 2000 design stratum numbers are not directly comparable, since the 1990 design contained some PSUs in New England and Hawaii that were based on minor civil divisions instead of counties, while the PSUs in the 2000 design are strictly county-based. The PSUs have also been redefined to correspond to the new Core-Based Statistical Area definitions and to improve efficiency in field operations.

[1] For detailed information on the 1990 sample redesign, see the Department of Labor, Bureau of Labor Statistics report, *Employment and Earnings,* Volume 41 Number 5, May 1994.

[2] The PSUs correspond to substate areas, counties, or groups of counties that are geographically contiguous.

Approximately 72,000 housing units were selected for sample from the mixed sampling frame in May. Based on eligibility criteria, 11 percent of these housing units were sent directly to Computer-Assisted Telephone Interviewing (CATI). The remaining units were assigned to interviewers for Computer-Assisted Personal Interviewing (CAPI)[3]. Of all housing units in sample, about 60,000 were determined to be eligible for interview. Interviewers obtained interviews at about 55,000 of these units. Noninterviews occur when the occupants are not found at home after repeated calls or are unavailable for some other reason.

May 2004 Supplement. In May 2004, in addition to the basic CPS questions, interviewers asked supplementary questions of all samplel households on work schedules and work at home.

Estimation Procedure. This survey's estimation procedure adjusts weighted sample results to agree with independently derived population estimates of the civilian noninstitutional population of the United States and states (including the District of Columbia). These population estimates, used as controls for the CPS, are prepared annually to agree with the most current set of population estimates that are released as part of the Census Bureau's population estimates and projections program.

The population controls for the nation are distributed by demographic characteristics in two ways:

- Age, sex, and race (White alone, Black alone, Asian alone, and all other groups combined), and
- Age, sex, and Hispanic origin.

The projections for the states are distributed by race (Black alone and all other race groups combined), age (0-15, 16-44, and 45 and over), and sex.

The independent estimates by age, sex, race, and Hispanic origin and for states by selected age groups and broad race categories are developed using the basic demographic accounting formula whereby the population from the latest decennial data is updated using data on the components of population change (births, deaths, and net international migration) with net internal migration as an additional component in the state population estimates.

The net international migration component in the population estimates includes a combination of:

- Legal migration to the United States,
- Emigration of foreign born and native people from the United States,
- Net movement between the United States and Puerto Rico,
- Estimates of temporary migration, and
- Estimates of net residual foreign-born population, which include unauthorized migration.

Because the latest available information on these components lag the survey date, it is necessary to make short-term projections of these components to develop the estimate for the survey date.

[3] For further information on CATI and CAPI and the eligibility criteria, please see: Technical Paper 63RV, *Current Population Survey: Design and Methodology*, U.S. Census Bureau, U.S. Department of Commerce, 2002. (http://www.census.gov/prod/2002pubs/tp63rv.pdf)

ACCURACY OF THE ESTIMATES

A sample survey estimate has two types of error: sampling and nonsampling. The accuracy of an estimate depends on both types of error. The nature of the sampling error is known given the survey design; the full extent of the nonsampling error is unknown.

Sampling Error. Since the CPS estimates come from a sample, they may differ from figures from an enumeration of the entire population using the same questionnaires, instructions, and enumerators. For a given estimator, the difference between an estimate based on a sample and the estimate that would result if the sample were to include the entire population is known as sampling error. Standard errors, as calculated by methods described in "Standard Errors and Their Use," are primarily measures of the magnitude of sampling error. However, they may include some nonsampling error.

Nonsampling Error. For a given estimator, the difference between the estimate that would result if the sample were to include the entire population and the true population value being estimated is known as nonsampling error. Sources of nonsampling errors include the following:

- Inability to get information about all sample cases (nonresponse)
- Definitional difficulties
- Differences in the interpretation of questions
- Respondent inability or unwillingness to provide correct information
- Respondent inability to recall information
- Errors made in data collection such as recording and coding data
- Errors made in processing the data
- Errors made in estimating values for missing data
- Failure to represent all units with the sample (undercoverage).

To minimize these errors, the Census Bureau employs quality control procedures in sample selection, wording of questions, interviewing, coding, data processing, and data analysis.

Two types of nonsampling error that can be examined to a limited extent are nonresponse and undercoverage.

Nonresponse. The effect of nonresponse cannot be measured directly, but one indication of its potential effect is the nonresponse rate. For the May 2004 basic CPS, the nonresponse rate was 7.8 percent. The nonresponse rate for the Work Schedules supplement was an additional 5.9 percent These two nonresponse rates lead to a combined supplement nonresponse rate of 13.2 percent.

Coverage. The concept of coverage in the survey sampling process is the extent to which the total population that could be selected for sample "covers" the survey's target population. CPS undercoverage results from missed housing units and missed persons within sample households. Overall CPS undercoverage for May 2004 is estimated to be about 11 percent. CPS undercoverage varies with age, sex, and race. Generally, undercoverage is larger for males than for females and larger for Blacks than for Non-Blacks.

The CPS weighting procedure partially corrects for bias due to undercoverage, but biases may still be present when people who are missed by the survey differ from those interviewed in ways other than age, race, sex, Hispanic ancestry, and state of residence. How this weighting procedure affects other variables in the survey is not precisely known. All of these considerations affect comparisons across different surveys or data sources.

A common measure of survey coverage is the coverage ratio, calculated as the estimated population before post-stratification divided by the independent population control. Table 1 shows May 2004 CPS coverage ratios for certain age-sex-race groups. The CPS coverage ratios can exhibit some variability from month to month.

Table 1. CPS Coverage Ratios: May 2004

	Totals			White Only		Black Only		Residual Race		Hispanic	
Age Group	**All People**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
0-15	0.90	0.91	0.90	0.93	0.91	0.81	0.78	0.92	0.98	0.93	0.93
16-19	0.88	0.88	0.88	0.91	0.90	0.77	0.74	0.83	0.94	1.00	0.93
20-24	0.79	0.78	0.80	0.80	0.82	0.63	0.76	0.77	0.65	0.77	0.84
25-34	0.83	0.80	0.86	0.83	0.87	0.68	0.80	0.75	0.82	0.76	0.89
35-44	0.88	0.85	0.91	0.87	0.94	0.70	0.78	0.88	0.90	0.84	0.91
45-54	0.93	0.92	0.94	0.93	0.96	0.83	0.85	0.88	0.92	0.83	0.93
55-64	0.94	0.93	0.94	0.95	0.95	0.84	0.89	0.91	0.88	0.92	0.95
65+	0.93	0.94	0.92	0.94	0.92	0.93	0.99	0.95	0.86	0.85	0.86
15+	0.89	0.87	0.90	0.89	0.92	0.76	0.83	0.85	0.86	0.83	0.90
0+	0.89	0.88	0.90	0.90	0.92	0.77	0.82	0.87	0.89	0.86	0.91

Notes: (1) The Residual Race group includes cases indicating a single race other than White or Black, and cases indicating two or more races.
(2) Hispanics may be of any race.

Comparability of Data. Data obtained from the CPS and other sources are not entirely comparable. This results from differences in interviewer training and experience and in differing survey processes. This is an example of nonsampling variability not reflected in the standard errors. Therefore, caution should be used when comparing results from different sources.

Caution should also be used when comparing the data from this microdata file, which reflects Census 2000-based controls, with microdata files from March 1994 through December 2001, which reflect 1990 census-based controls. Caution should also be used when comparing the data from this microdata file to certain microdata files from 2002, namely June, October, and November, which contain both Census 2000-based estimates and 1990 census-based estimates. When comparing estimates, the same controls should be used when possible. Microdata files from previous years reflect the latest available census-based controls. Although this change in population controls had relatively little impact on summary measures such as averages, medians, and percentage distributions, it did have a significant impact on levels. For example, use of Census 2000-based controls results in about a one percent increase from the 1990 census-based controls in the civilian noninstitutional population and in the number of families and households. Thus, estimates of levels for data collected 2002 and later years will differ from those for earlier years by more than what could be attributed to actual changes in the population. These differences could be disproportionately greater for certain subpopulation groups than for the total population.

Users should also exercise caution due to changes caused by the phase-in of the Census 2000 files. During this time period, CPS data are collected from sample designs based on different censuses. Three features of the new CPS design have the potential of affecting published estimates: (1) the temporary disruption of the rotation pattern from August 2004 through June 2005 for a comparatively small portion of the sample (which doesn't affect this survey), (2) the change in sample areas, and (3) the introduction of the new Core-Based Statistical Areas (formerly called metropolitan areas). Most of the known effect on estimates during and after the sample redesign will be the result of changing from 1990 to 2000 geographic definitions. Research has shown that the national-level estimates of the metropolitan and nonmetropolitan populations should not change appreciably because of the new sample design. However, users should still exercise caution when comparing metropolitan and nonmetropolitan estimates across years with a design change, especially at the state level.

Caution should also be used when comparing Hispanic estimates over time. No independent population control totals for people of Hispanic ancestry were used before 1985.

A Nonsampling Error Warning. Since the full extent of the nonsampling error is unknown, one should be particularly careful when interpreting results based on small differences between estimates. Even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test. Caution should also be used when interpreting results based on a relatively small number of cases. Summary measures probably do not reveal useful information when computed on a subpopulation smaller than 75,000.

For additional information on nonsampling error including the possible impact on CPS data when known, refer to

- Statistical Policy Working Paper 3, *An Error Profile: Employment as Measured by the Current Population Survey*, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce 1978. (http://www.fcsm.gov/working-papers/spp.html)

- Technical Paper 63RV, *The Current Population Survey: Design and Methodology*, U.S. Census Bureau, U.S. Department of Commerce, 2002. (http://www.census.gov/prod/2002pubs/tp63rv.pdf)

Standard Errors and Their Use. The sample estimate and its standard error enable one to construct a confidence interval. A confidence interval is a range that would include the average result of all possible samples with a known probability. For example, if all possible samples were surveyed under essentially the same general conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then approximately 90 percent of the intervals from 1.645 standard errors below the estimate to 1.645 standard errors above the estimate would include the average result of all possible samples.

A particular confidence interval may or may not contain the average estimate derived from all possible samples. However, one can say with specified confidence that the interval includes the average estimate calculated from all possible samples.

Standard errors may also be used to perform hypothesis testing. This is a procedure for distinguishing between population parameters using sample estimates. The most common type of hypothesis is that the population parameters are different. An example of this would be comparing the percentage of men who were part-time workers to the percentage of women who were part-time workers.

Tests may be performed at various levels of significance. A significance level is the probability of concluding that the characteristics are different when, in fact, they are the same. For example, to conclude that two characteristics are different at the 0.10 level of significance, the absolute value of the estimated difference between characteristics must be greater than or equal to 1.645 times the standard error of the difference.

The Census Bureau uses 90-percent confidence intervals and 0.10 levels of significance to determine statistical validity. Consult standard statistical textbooks for alternative criteria.

Estimating Standard Errors. The Census Bureau uses replication methods to estimate the standard errors of CPS estimates. These methods primarily measure the magnitude of sampling error. However, they do measure some effects of nonsampling error as well. They do not measure systematic biases in the data due to nonsampling error. Bias is the average over all possible samples of the differences between the sample estimates and the true value.

Generalized Variance Parameters. While it is possible to compute and present an estimate of the standard error based on the survey data for each estimate in a report, there are a number of reasons why this is not done. A presentation of the individual standard errors would be of limited use, since one could not possibly predict all of the combinations of results that may be of interest to data users. Additionally, variance estimates are based on sample data and have variances of their own. Therefore, some method of stabilizing these estimates of variance, for example, by generalizing or averaging over time, may be used to improve their reliability.

Experience has shown that certain groups of estimates have similar relationships between their variances and expected values. Modeling or generalizing may provide more stable variance estimates by taking advantage of these similarities. The generalized variance function is a simple model that expresses the variance as a function of the expected value of the survey estimate. The parameters of the generalized variance function are estimated using direct replicate variances. These generalized variance parameters provide a relatively easy method to obtain approximate standard errors for numerous characteristics. In this source and accuracy statement, Table 2 provides the generalized variance parameters for labor force estimates.

Standard Errors of Estimated Numbers. The approximate standard error, s_x, of an estimated number from this microdata file can be obtained by using this formula:

$$s_x = \sqrt{ax^2 + bx} \qquad (1)$$

Here x is the size of the estimate and a and b are the parameters in Table 2 associated with the particular type of characteristic.

When calculating standard errors from cross-tabulations involving different characteristics, use the set of parameters for the characteristic that will give the largest standard error.

For information on calculating standard errors for labor force data from the CPS which involve quarterly or yearly averages, see “Explanatory Notes and Estimates of Error: Household Data” in *Employment and Earnings*, a monthly report published by the U.S. Bureau of Labor Statistics.

Illustration No. 1

Suppose there were 4,292,000 unemployed men (ages 16 and up) in the civilian labor force[4]. Use the appropriate parameters from Table 2 and Formula (1) to get:

Illustration 1	
Number of unemployed men in the civilian labor force (x)	4,292,000
a parameter (a)	-0.000035
b parameter (b)	2,927
Standard error	109,000
90% confidence interval	4,113,000 to 4,471,000

The standard error is calculated as

$$s_x = \sqrt{-0.000035 \times 4{,}292{,}000^2 + 2{,}927 \times 4{,}292{,}000} = 109{,}000$$

The 90-percent confidence interval is calculated as 4,292,000 ± 1.645 × 109,000.

A conclusion that the average estimate derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all possible samples.

Standard Errors of Estimated Percentages. The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends on both the size of the percentage and its base. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more. When the numerator and denominator of the percentage are in different categories, use the parameter from Table 2 as indicated by the numerator.

The approximate standard error, $s_{x,p}$, of an estimated percentage can be obtained by using the formula:

$$s_{x,p} = \sqrt{\frac{b}{x} p(100 - p)} \qquad (2)$$

Here x is the total number of people, families, households, or unrelated individuals in the base of the percentage, p is the percentage ($0 \leq p \leq 100$), and b is the parameter in Table 2 associated with the characteristic in the numerator of the percentage.

[4] The CPS collects labor force participation data on all respondents of age 15 and over. However, the BLS defines the civilian labor force to include only persons of age 16 and over. This example’s counts are consistent with the BLS definition.

Illustration No. 2

Suppose that of 139,164,000 workers, 48,820,000, or 35.1 percent, were on flexible schedules. Use the appropriate parameter from Table 2 and Formula (2) to get

Illustration 2	
Percentage of workers who were on flexible schedules (p)	35.1
Base (x)	139,164,000
b parameter (b)	1,586
Standard error	0.16
90% confidence interval	34.8 to 35.4

The standard error is calculated as

$$s_{x,p} = \sqrt{\frac{1{,}586}{139{,}164{,}000} \times 35.1 \times (100 - 35.1)} = 0.16$$

The 90-percent confidence interval for the percentage of workers on flexible schedules is calculated as 35.1 ± 1.645 × 0.16.

Standard Errors of Differences. The standard error of the difference between two sample estimates is approximately equal to

$$s_{x-y} = \sqrt{s_x^2 + s_y^2} \tag{3}$$

where s_x and s_y are the standard errors of the estimates, x and y. The estimates can be numbers, percentages, ratios, etc. This will represent the actual standard error quite accurately for the difference between estimates of the same characteristic in two different areas, or for the difference between separate and uncorrelated characteristics in the same area. However, if there is a high positive (negative) correlation between the two characteristics, the formula will overestimate (underestimate) the true standard error.

For information on calculating standard errors for labor force data from the CPS which involve differences in consecutive quarterly or yearly averages, consecutive month-to-month differences in estimates, and consecutive year-to-year differences in monthly estimates see “Explanatory Notes and Estimates of Error: Household Data” in *Employment and Earnings*, a monthly report published by the U.S. Bureau of Labor Statistics.

Illustration No. 3

Suppose that of 7,110,000 employed men between 20 and 24 years of age, 1,522,000 or 21.4 percent were part-time workers, and of the 6,437,000 employed women between 20 and 24 years of age, 2,161,000 or 33.6 percent were part-time workers. Use the appropriate parameters from Table 2 and formulas (2) and (3) to get

Illustration 3			
	Men (x)	Women (y)	Difference
Percentage between 20-24 who were part-time workers (p)	21.4	33.6	12.2
Base (x)	7,110,000	6,437,000	-
b parameter (b)	2,927	2,693	-
Standard error	0.83	0.97	1.28
90% confidence interval	20.0 to 22.8	32.0 to 35.2	10.1 to 14.3

The standard error of the difference is calculated as

$$s_{x-y} = \sqrt{0.83^2 + 0.97^2} = 1.28$$

The 90-percent confidence interval around the difference is calculated as 12.2 ± 1.645 × 1.28. Since this interval does not include zero, we can conclude with 90 percent confidence that the percentage of part-time women workers between 20-24 years of age is greater than the percentage of part-time men workers between 20-24 years of age.

Table 2. Parameters for Computation of Standard Errors for Labor Force Characteristics: May 2004

Characteristic	a	b
Labor Force and Not in Labor Force Data Other than Agricultural Employment and Unemployment		
Total or White	-0.000008	1,586
Men	-0.000035	2,927
Women	-0.000033	2,693
Both sexes, 16 to 19 years	-0.000244	3,005
Black	-0.000154	3,296
Men	-0.000336	3,332
Women	-0.000282	2,944
Both sexes, 16 to 19 years	-0.001531	3,296
Hispanic Ancestry	-0.000187	3,296
Men	-0.000363	3,332
Women	-0.000380	2,944
Both sexes, 16 to 19 years	-0.001822	3,296
Asian and Pacific Islander (API)	-0.000272	2,749
Men	-0.000569	2,749
Women	-0.000521	2,749
Unemployment		
Total or White	-0.000017	3,005
Men	-0.000035	2,927
Women	-0.000033	2,693
Both sexes, 16 to 19 years	-0.000244	3,005
Black	-0.000154	3,296
Men	-0.000336	3,332
Women	-0.000282	2,944
Both sexes, 16 to 19 years	-0.001531	3,296
Hispanic Ancestry	-0.000187	3,296
Men	-0.000363	3,332
Women	-0.000380	2,944
Both sexes, 16 to 19 years	-0.001822	3,296
Asian and Pacific Islander (API)	-0.000272	2,749
Men	-0.000569	2,749
Women	-0.000521	2,749
Agricultural Employment		
Total	0.001345	2,989

NOTE: (1) These parameters are to be applied to basic CPS monthly labor force estimates.

(2) For foreign-born and noncitizen characteristics for Total and White, the a and b parameters should be multiplied by 1.3. No adjustment is necessary for foreign-born and noncitizen characteristics for Blacks, APIs, and Hispanics.