Source and Accuracy Statement for School Enrollment Estimates from the October 2002 CPS Microdata File

Table of Contents

Tables

Source and Accuracy of Estimates for the October 2002 CPS Microdata File on School Enrollment

SOURCE OF DATA

The data in this microdata file come from the October 2002 Current Population Survey (CPS). The Census Bureau conducts the survey every month, although this file has only October data. The October survey uses two sets of questions, the basic CPS and the supplement.

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

The monthly CPS sample is a multistage probability sample with coverage in all 50 states and the District of Columbia. The sample was selected from 1990 Decennial Census files and is continually updated to account for new residential construction. To obtain the sample, the United States was divided into 2,007 geographic areas. In most states, a geographic area consisted of a county or several contiguous counties. In some areas of New England and Hawaii, minor civil divisions are used instead of counties. These 2,007 geographic areas were then grouped into 754 strata, and one geographic area was selected from each stratum.

About 60,000 occupied households are eligible for interview every month out of the 754 strata. Interviewers are unable to obtain interviews at about 4,500 of these units. This occurs when the occupants are not found at home after repeated calls or are unavailable for some other reason. The number of households that are eligible for interview in the basic CPS increased from 50,000 to 60,000 in July of 2001. With the increase in eligible households, the number of units where interviewers were unable to obtain an interview increased from 3,200 to 4,500.

October Supplement. In addition to the basic CPS questions, interviewers asked supplementary questions in October about school enrollment for all household members three years old and over.

Sample Redesign. Since the introduction of the CPS, the Census Bureau has redesigned the CPS sample several times. These redesigns have improved the quality and accuracy of the data and have satisfied changing data needs. The most recent changes were phased in and implementation was completed in July 1995.

Estimation Procedure. This survey's estimation procedure adjusts weighted sample results to agree with independent estimates of the civilian noninstitutional population of the United States by age, sex, race, Hispanic/non-Hispanic ancestry, and state of residence. The adjusted estimate is called the post-stratification ratio estimate. The independent estimates are calculated based on information from three primary sources:

- The 2000 Decennial Census of Population and Housing
- Statistics on births, deaths, immigration, and emigration
- Statistics on the size of the Armed Forces

The independent population estimates include some, but not all, unauthorized migrants.

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, however, is unknown.

Sampling Error. Since the CPS estimates come from a sample, they may differ from figures from a complete census using the same questionnaires, instructions, and enumerators. This possible variation in the estimates due to sampling error is known as "sampling variability." Standard errors, as calculated by methods described in "Standard Errors and Their Use" are primarily measures of sampling variability, although they may include some nonsampling error.

Nonsampling Error. All other sources of error in the survey estimates are collectively called nonsampling error. Sources of nonsampling error include the following:

- Inability to obtain 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).

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 October 2002 basic CPS, the nonresponse rate was 6.7%. The nonresponse rate for the October supplement was an additional 6.2%. These two nonresponse rates lead to a total nonresponse rate of 12.5%.

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 people within sample households. Overall CPS undercoverage is estimated to be about 10 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 Current Population Survey weighting procedure uses ratio estimation whereby sample estimates are adjusted to independent estimates of the national population by age, race, sex and

Hispanic ancestry. This weighting 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, and Hispanic ancestry. 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, the estimated population before post-stratification divided by the independent population control. Table 1 shows October 2002 CPS coverage ratios for certain age-sex-race groups. The CPS coverage ratios can exhibit some variability from month to month. Other Census Bureau household surveys experience similar coverage.

Table 1. CPS Coverage Ratios for October 2002

	Non-Black		Black		All Persons		
Age	M	F	M	F	M	F	Total
0-14	0.94	0.92	0.85	0.88	0.92	0.92	0.92
15	0.94	0.97	0.83	0.97	0.92	0.97	0.94
16-19	0.89	0.91	0.81	0.77	0.88	0.89	0.88
20-29	0.79	0.84	0.67	0.78	0.78	0.83	0.80
30-39	0.87	0.91	0.76	0.90	0.86	0.91	0.88
40-49	0.91	0.94	0.87	0.93	0.91	0.93	0.92
50-59	0.94	0.93	0.90	0.91	0.93	0.93	0.93
60-64	0.88	0.91	0.94	0.94	0.89	0.92	0.90
65-69	0.95	0.94	0.96	0.92	0.96	0.94	0.95
70+	0.92	0.92	0.96	0.92	0.92	0.92	0.92
15+	0.89	0.91	0.81	0.88	0.88	0.91	0.89
0+	0.90	0.91	0.83	0.88	0.89	0.91	0.90

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.

A number of changes were made in data collection and estimation procedures beginning with the January 1994 CPS. The major change was the use of a new questionnaire. The questionnaire was redesigned to measure the official labor force concepts more precisely, to expand the amount of data available, to implement several definitional changes, and to adapt to a computer-assisted interviewing environment. The supplemental questions were also modified for adaptation to computer-assisted interviewing, although there were no changes in definitions and concepts. See Appendix C of Report P-60, No. 188 on "Conversion to a Computer Assisted Questionnaire" for a description of these changes and the effect they had on the data. Due to

these and other changes, one should use caution when comparing estimates from data collected in 1994 and later years with estimates from earlier years.

Caution should also be used when comparing data from this microdata file, which reflects 2000 census-based population controls, with microdata files from October 1994-2000, which reflect 1990 census-based population controls. Microdata files from previous years reflect the latest available census-based population controls.

Although the change in population controls had relatively little impact on summary measures such as means, medians, and percentage distributions, it did have a significant impact on levels. For example, use of 2000 census-based population controls results in about a one-percent increase in the civilian noninstitutional population and in the number of families and households. Thus, estimates of levels for data collected in 2001 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.

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.

Based on the results of each decennial census, the Census Bureau gradually introduces a new sample design for the CPS[1]. During this phase-in period, CPS data are collected from sample designs based on different censuses. While most CPS estimates were unaffected by this mixed sample, geographic estimates are subject to greater error and variability. Users should exercise caution when comparing estimates across years for metropolitan/ nonmetropolitan categories.

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 base[2] smaller than 75,000.

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

- 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.

---

[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] subpopulation

(http://www.fcsm.gov/working-papers/spp.html)

- 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)

Standard Errors and Their Use. A number of approximations are required to derive, at a moderate cost, standard errors applicable to all the estimates in this microdata file. Instead of providing an individual standard error for each estimate, parameters are provided to calculate standard errors for various types of characteristics. These parameters are listed in Tables 2 and 3. Also, tables are provided that allow the calculation of parameters for prior years and parameters for U.S. regions. Tables 4 and 4A provide factors to derive prior year parameters. Table 5 provides factors to derive U.S. regional parameters.

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 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, a procedure for distinguishing between population parameters using sample estimates. One common type of hypothesis is that the population parameters are different. An example of this would be comparing the percentage of employed males 20 to 24 years old working part time to the percentage of employed females in the same age group who were part-time workers. An illustration of this is included in the following pages.

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. To conclude that two parameters 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. To estimate the standard error of a CPS estimate, the Census Bureau uses replicated variance estimation methods. 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. It is possible to compute and present an estimate of the standard error based on the survey data for each estimate in a report, but 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, is needed to improve their reliability.

Experience has shown that certain groups of estimates have a similar relationship between their variance and expected value. Modeling or generalization 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 model are estimated using direct replicate variances. These generalized variance parameters provide a relatively easy method to obtain approximate standard errors for numerous characteristics. Table 2 provides the labor force parameters, table 3 provides the school enrollment parameters, and tables 4, 4A, and 5 provide factors for use with the parameters.

Standard Errors of Estimated Numbers. The approximate standard error, s_x, of an estimated number, **with the exception of school enrollment estimates**, from this microdata file can be obtained using the following 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 which will give the largest standard error.

Illustration No. 1
In October 2002, suppose there were 3,900,000 unemployed men in the civilian labor force. Use the appropriate parameters from Table 2 and formula (1) to get the following:

Number, x	3,900,000
a parameter	-0.000035
b parameter	2,927
Standard error	104,000
90% conf. int.	3,729,000 to 4,071,000

The standard error is calculated as follows:

$$s_x = \sqrt{-0.000035 \times 3{,}900{,}000^2 + 2{,}927 \times 3{,}900{,}000} = 104{,}000$$

The 90-percent confidence interval is calculated as 3,900,000 ± 1.645 × 104,000.

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 School Enrollment Numbers. The approximate standard error, s_x, of an estimated school enrollment number from this microdata file can be obtained using the following formula:

$$s_x = \sqrt{-\left(\frac{b}{T}\right)x^2 + bx} \qquad (2)$$

Here x is the size of the estimate, T is the total number of persons in a specific age group and b is the parameter in Table 3 associated with the particular type of characteristic. If T is not known, for Total or White use 100,000,000; for Blacks and Hispanic use 10,000,000. When calculating standard errors for numbers from cross-tabulations involving different characteristics, use the set of parameters for the characteristic which will give the largest standard error.

Illustration No. 2
Suppose there were 4,100,000 three and four year olds enrolled in school and 7,900,000 children in that age group in October 2002. Use the appropriate b parameter from Table 3 and formula (2) to get the following:

Number, x	4,100,000
Total, T	7,900,000
b parameter	2,453
Standard error	70,000
90% conf. int.	3,985,000 to 4,215,000

The standard error is calculated as follows:

$$s_x = \sqrt{-\frac{2{,}453}{7{,}900{,}000} \times 4{,}100{,}000^2 + 2{,}453 \times 4{,}100{,}000} = 70{,}000$$

The 90-percent confidence interval for this estimate is approximately 3,985,000 to 4,215,000 (i.e., 4,100,000 $\pm$ 1.645 × 70,000). Therefore, 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 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 or 3 indicated by the numerator.

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

$$s_{x,p} = \sqrt{\frac{b}{x} p(100 - p)} \qquad (3)$$

Here x is the total number of persons, families, households, or unrelated individuals in the base of the percentage, p is the percentage ($0 \le p \le 100$), and b is the parameter in Table 2 or 3 associated with the characteristic in the numerator of the percentage.

Illustration No. 3
In October 2002, suppose there were 16,000,000 persons aged 18 to 21, and 44.0 percent were enrolled in college. Use the appropriate parameter from Table 3 and formula (3) to get the following:

Percentage, p	44.0
Base, x	16,000,000
b parameter	2,131
Standard error	0.57
90% conf. int.	43.06 to 44.94

The standard error is calculated as follows:

$$s_{x,p} = \sqrt{\frac{2{,}131}{16{,}000{,}000} \times 44.00 \times (100.0 - 44.00)} = 0.57$$

The 90 percent confidence interval for the estimated percentage of persons aged 18 to 21 in October 2002 enrolled in college is from 43.06 to 44.94 percent (i.e., $44.00 \pm 1.645 \times 0.57$).

Standard Error of a Difference. The standard error of the difference between two sample estimates is approximately equal to the following:

$$s_{x-y} = \sqrt{s_x^2 + s_y^2} \tag{4}$$

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 result in accurate estimates of the standard error 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.

Illustration No. 4
Suppose that of the 6,850,000 employed men between 20-24 years of age in October 2002, 20.8 percent were part-time workers, and of the 6,400,000 employed women between 20-24 years of age, 34.6 percent were part-time workers. Use the appropriate parameters from Table 2 and formulas (3) and (4) to get the following:

	x	y	difference
Percentage, p	20.8	34.6	13.8
Number, x	6,850,000	6,400,000	-
b parameter	2,927	2,693	-
Standard error	0.84	0.98	1.29
90% conf. int.	19.42 to 22.18	32.99 to 36.21	11.68 to 15.92

The standard error of the difference is calculated as follows:

$$s_{x-y} = \sqrt{0.84^2 + 0.98^2} = 1.29$$

The 90-percent confidence interval around the difference is calculated as $13.8 \pm 1.645 \times 1.29$. 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: October 2002

Characteristic	a	b
Civilian Labor Force, Employed, and Not in Labor Force		
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
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
Agricultural Employment	0.001345	2,989

Notes: These parameters are to be applied to basic CPS monthly labor force estimates. 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 and Hispanics.

Table 3. Parameters for Computation of Standard Errors for School Enrollment Characteristics: October 2002

Characteristics	Total or White b	Black b	Hispanic b
People			
Persons Enrolled in School:			
Total	2,131	2,410	2,744
Children 13 and under	2,453	2,775	3,159
Marital Status, Household and Family Characteristics, Health Insurance			
Some household members	4,687	6,733	11,347
All household members	5,695	9,929	16,733
Families, Households, or Unrelated Individuals			
Income, Earnings	2,016	2,201	3,709
Marital Status, Household and Family Characteristics, Educational Attainment, Population by Age and/or Sex	1,860	1,683	2,836

Notes: The b parameters should be multiplied by 1.5 for nonmetropolitan residence categories. The b parameters should be multiplied by the factors in Table 5 for regional data.

In 1994, we calculated school enrollment parameters directly from the 1994 CPS data. Since that time, the school enrollment parameters have been based on these updated parameters. Therefore, when calculating past school enrollment parameters, a separate set of year factors should be used.

Table 4 shows the prior year factors to apply to the **Non-School Enrollment** parameters while Table 4A shows prior year factors to apply to **School Enrollment** parameters.

Table 4. Year Factors for Non-School Enrollment Characteristics (1942-2001)

Time Period	Total/White	Black	Hispanic
July 2001 - Present	1.00	1.00	1.00
January 1996 - June 2001	1.11	1.11	1.11
April 1989 - December 1995	1.03	1.03	1.03
April 1988 - March 1989	1.14	1.14	1.20
January 1985 - March 1988	0.96	0.96	0.96
January 1982 - December 1984	0.96	0.96	1.35
March 1973 - December 1981	0.86	0.86	1.20
January 1967 - February 1973	0.86	0.86	1.20
May 1956 - December 1966	1.29	1.29	1.81
August 1942 - April 1956	1.93	1.93	2.71

Note: These factors are for use with the 2002 non-School Enrollment 'a' and 'b' parameters.

Table 4A. Year Factors for School Enrollment Characteristics (1945-2001)

Time Period	Total or White	Black	Hispanic
July 2001 - Present	1.00	1.00	1.00
January 1996 - June 2001	1.11	1.11	1.11
March 1994 - December 1995	1.03	1.03	1.03
April 1989 - February 1994	1.19	1.42	2.10
April 1988 - March 1989	1.32	1.58	2.45
January 1985 - March 1988	1.11	1.33	1.97
January 1982 - December 1984	1.11	1.33	2.76
March 1973 - December 1981	0.99	1.19	2.46
January 1967 - February 1973	0.99	1.19	2.46
May 1956 - December 1966	1.49	1.78	3.69
October 1945 - April 1956	2.24	2.67	5.54

Note: These factors are for use with the 2002 School Enrollment 'a' and 'b' parameters.

Table 5 provides the U.S. regional factors to apply to parameters in order to calculate standard errors for U.S. regional estimates.

Table 5. Regional Factors to Apply to 2002 Parameters

Type of Characteristic	factor
U. S. Totals:	1.00
Regions:	
Northeast	0.90
Midwest	0.93
South	1.14
West	1.14