Accuracy of the 2005-2007 ACS 3-Year Data

INTRODUCTION

The data contained in these data products are based on the American Community Survey (ACS) and Puerto Rico Community Survey (PRCS) sample interviewed from January 1, 2005 through December 31, 2007. Data products were produced for three sets of 1-year estimates (2005, 2006 and 2007), in addition to this set of 3-year estimates. In 2010, the Census Bureau plans to publish the first 5-year data products, based on data collected in 2005 through 2009. Unless otherwise specified, the term "ACS" in this document will refer to both the ACS and PRCS.

In general, ACS estimates are period estimates that describe the average characteristics of population and housing over a period of data collection. The 2005-2007 estimates are averages over the period from January 1, 2005 to December 31, 2007. Multiyear estimates cannot be used to say what is going on in any particular year in the period, only what the average value is over the full period.

The ACS sample is selected from all counties and county-equivalents in the United States, and all municipios in Puerto Rico (PR). In 2006, the ACS began collection of data from sampled persons in group quarters (GQs) – for example, military barracks, college dormitories, nursing homes, and correctional facilities. Persons in group quarters are included with persons in housing units (HUs) in all 2005-2007 ACS estimates based on the total population.

The ACS, like any other statistical activity, is subject to error. The purpose of this documentation is to provide data users with a basic understanding of the ACS sample design, estimation methodology, and accuracy of the 2005-2007 ACS data. The ACS is sponsored by the U.S. Census Bureau, and is part of the 2010 Decennial Census Program.

DATA COLLECTION

The ACS employs three modes of data collection:

- Mailout/Mailback
- Computer Assisted Telephone Interview (CATI)
- Computer Assisted Personal Interview (CAPI)

The general timing of data collection is:

Month 1:	Addresses determined to be mailable are sent a questionnaire via the U.S. Postal Service.
Month 2:	All mail non-responding addresses with an available phone number are sent to CATI.
Month 3:	A sample of mail non-responses without a phone number, CATI non-responses, and unmailable addresses are selected and sent to CAPI.

SAMPLE DESIGN

Sampling rates are assigned independently at the census block level. A measure of size is calculated for each of the following governmental units (GUs):

- Counties
- Places (active, functioning governmental units)
- School Districts (elementary, secondary, and unified)
- American Indian Areas (included Tribal Subdivisions beginning in 2007)
- Minor Civil Divisions (MCDs) – in Massachusetts, New York, Pennsylvania, and Wisconsin (these are the states where MCDs are active, functioning governmental units)

Each block is then assigned the smallest measure of size from the set of all governmental units it is a part of (GUMOS).

The measure of size for all geographic entities for all areas except American Indian Areas, is an estimate of the number of occupied housing units in the area. This was calculated by multiplying the number of ACS addresses by the occupancy rate from Census 2000 at the block level. For American Indian Areas the measure of size is the estimated number of occupied HUs multiplied by the proportion of people reporting American Indian (alone or in combination) in Census 2000. A measure of size for each census tract (TRACTMOS) was also calculated in the same manner.

Table 2. Sampling Rates 2005-2007 for the United States and Puerto Rico

Sampling Rate Category	Sampling Rates	
	United States	Puerto Rico
Blocks in smallest governmental units (GUMOS < 200)	30.0%	30.0%
Blocks in smaller governmental units (200 ≤ GUMOS < 800)	20.4%	24.3%
Blocks in small governmental units (800 ≤ GUMOS ≤ 1200)	10.2%	12.1%
Blocks in large tracts (GUMOS > 1200, TRACTMOS ≥ 2000) where mailable addresses ≥ 75% and predicted levels of completed mail and CATI interviews prior to CAPI subsampling > 60%	4.6%	6.0%
Other Blocks in large tracts (GUMOS > 1200, TRACTMOS ≥ 2000)	5.0%	
All other blocks (GUMOS > 1200, TRACTMOS < 2000) where mailable addresses ≥ 75% and predicted levels of completed mail and CATI interviews prior to CAPI subsampling > 60%	6.2%	8.1%
All other blocks (GUMOS > 1200, TRACTMOS < 2000)	6.8%	

All addresses determined to be unmailable are subsampled for the CAPI phase of data collection at a rate of 2-in-3. Unmailable addresses do not go to the CATI phase of data collection. Subsequent to CATI, all addresses for which no response has been obtained prior to CAPI are subsampled. Beginning with the CAPI for the January 2005 panel (March 2005 data collection), the CAPI subsampling rate was based on the expected rate of completed mail and CATI interviews at the tract level.

Table 3. CAPI Subsampling Rates 2005-2007 for the United States and Puerto Rico

Address and Tract Characteristics	CAPI Subsampling Rates
United States	
Unmailable addresses and addresses in Remote Alaska	66.7%
Mailable addresses in tracts with predicted levels of completed mail and CATI interviews prior to CAPI subsampling between 0% and less than 36%	50%
Mailable addresses in tracts with predicted levels of completed mail and CATI interviews prior to CAPI subsampling greater than 35% and less than 51%	40%
Mailable addresses in other tracts	33.3%
Puerto Rico	
Unmailable addresses	66.7%
Mailable addresses	50%

For a more detailed description of the ACS sampling methodology, see the 2007 Accuracy of the Data document (http://www.census.gov/acs/www/Downloads/ACS/accuracy2007.pdf). For more information relating to sampling in a specific year, please refer to the individual year's Accuracy of the Data document http://www.census.gov/acs/www/UseData/Accuracy/Accuracy1.htm.

WEIGHTING METHODOLOGY

The multiyear estimates should be interpreted as estimates that describe a time period rather a specific reference year. For example, a 3-year estimate for the poverty rate of a given area describes the total set of people who lived in that area over those three years much the same way as a 1-year estimate for the same characteristic describes the set of people who lived in that area over one year. The only fundamental difference between the estimates is the number of months of collected data which are considered in forming the estimate. For this reason, the estimation procedure used for the multiyear estimates is an extension of the 2007 1-year estimation procedure. In this document only the procedures that are unique to the multiyear estimates are discussed.

To weight the 3-year estimates, 36 months of collected data are pooled together. The exception is for group quarters data. The ACS did not begin full implementation for group quarters until 2006. Thus, the 2005-2007 group quarters estimates use only data collected in the years 2006 and 2007.

The pooled data are then reweighted using the procedures developed for the 2007 1-year estimates with a few adjustments. These adjustments concern geography, month-specific weighting steps, population and housing unit controls, and inflation factors. In addition, a new step has been added to the multiyear estimation process.

For the 1-year estimation, the tabulation geography for the data is based on the boundaries defined on January 1 of the tabulation year, which is consistent with the tabulation geography used to produce the population estimates. All sample addresses are updated with this geography prior to weighting. For the multiyear estimation, the tabulation geography for the data is referenced to the final year in the multiyear period. For example, the 2005-2007 period uses the 2007 reference geography. Thus, all data collected over the period of 2005-2007 in the blocks that are contained in the 2007 boundaries for a given place are tabulated as though they were a part of that place for the entire period.

Some of the weighting steps use the month of tabulation in forming the weighting cells within which the weighting adjustments are made. One such example is the non-interview adjustment. In these weighting steps, the month of tabulation is used independent of year. Thus, sample cases from May 2005, May 2006, and May 2007 are combined into one weighting cell.

Since the multiyear estimates represent estimates for the period, the controls are not a single year's housing or population estimates from the Population Estimates Program, but rather are an average of these estimates over the period. For the housing unit controls, a simple average of the 1-year housing unit estimates over the period is calculated for each county. The version or vintage of estimates used is always the last year of the period since these are considered to be the most up-to-date and are created using a consistent methodology. For example, the housing unit control used for a given county in the 2005-2007 weighting is equal to the simple average of the 2005, 2006, and 2007 estimates that were produced using the 2007 methodology (the 2007 vintage). Likewise, the population controls by race, ethnicity, age, and sex are obtained by taking a simple average of the 1-year population estimates of the county by race, ethnicity, age, and sex. For example, the 2005-2007 control total used for Hispanic males age 20-24 in a given county would be obtained by averaging the 1-year estimates for that demographic group for 2005, 2006, and 2007.

Monetary values for the ACS 3-year estimates are inflation-adjusted to the final year of the period. For example, the 2005-2007 ACS 3-year estimates are tabulated using 2007-adjusted dollars. These adjustments use the national Consumer Price Index (CPI) since a regional-based CPI is not available for the entire country.

The new multiyear specific step is a model-assisted (generalized regression or GREG) weighting step. The objective of this additional step is to reduce the variances of base demographics at the place and MCD level in the 3-year estimates. While reducing the variances, the estimates themselves are relatively unchanged. This process involves linking administrative record data with ACS data.

For a more detailed description of the ACS estimation methodology, see the 2007 Accuracy of the Data document (http://www.census.gov/acs/www/Downloads/ACS/accuracy2007.pdf). For more information relating to estimation in a specific year, please refer to that individual year's

Accuracy of the Data document
(http://www.census.gov/acs/www/UseData/Accuracy/Accuracy1.htm).

CONFIDENTIALITY OF THE DATA

The Census Bureau has modified or suppressed some data on this site to protect confidentiality. Title 13 United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's data can be identified.

The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed.

- Title 13, United States Code: Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 be maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to five years in prison and up to $250,000 in fines for wrongful disclosure of confidential census information.

- Disclosure Limitation: Disclosure limitation is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual that has provided information under a pledge of confidentiality. For data tabulations the Census Bureau uses disclosure limitation procedures to modify or remove the characteristics that put confidential information at risk for disclosure. Although it may appear that a table shows information about a specific individual, the Census Bureau has taken steps to disguise or suppress the original data while making sure the results are still useful. The techniques used by the Census Bureau to protect confidentiality in tabulations vary, depending on the type of data.

- Data Swapping: Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases when creating a table. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics (such as the same number of adults and same number of children). Because the swap often occurs within a neighboring area, there is no effect on the marginal totals for the area or for totals that include data from multiple areas. Because of data swapping, users should not assume that tables with cells having a value of one or two reveal information about specific individuals. Data swapping procedures were first used in the 1990 Census, and were used again in Census 2000.

The data use the same disclosure limitation methodology as the original 1-year data. The confidentiality edit was previously applied to the raw data files when they were created to produce the 1-year estimates and these same data files with the original confidentiality edit were used to produce the 3-year and 5-year estimates.

ERRORS IN THE DATA

- Sampling Error — The data in the ACS products are estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology. The estimates from the chosen sample also differ from other samples of housing units and persons within those housing units. Sampling error in data arises due to the use of probability sampling, which is necessary to ensure the integrity and representativeness of sample survey results. The implementation of statistical sampling procedures provides the basis for the statistical analysis of sample data.

- Nonsampling Error — In addition to sampling error, data users should realize that other types of errors may be introduced during any of the various complex operations used to collect and process survey data. For example, operations such as data entry from questionnaires and editing may introduce error into the estimates. Another source is through the use of controls in the weighting. The controls are designed to mitigate the effects of systematic undercoverage of certain groups who are difficult to enumerate and to reduce the variance. The controls are based on the population estimates extrapolated from the previous census. Errors can be brought into the data if the extrapolation methods do not properly reflect the population. However, the potential risk from using the controls in the weighting process is offset by far greater benefits to the ACS estimates. These benefits include reducing the effects of a larger coverage problem found in most surveys, including the ACS, and the reduction of standard errors of ACS estimates. These and other sources of error contribute to the nonsampling error component of the total error of survey estimates. Nonsampling errors may affect the data in two ways. Errors that are introduced randomly increase the variability of the data. Systematic errors which are consistent in one direction introduce bias into the results of a sample survey. The Census Bureau protects against the effect of systematic errors on survey estimates by conducting extensive research and evaluation programs on sampling techniques, questionnaire design, and data collection and processing procedures. In addition, an important goal of the ACS is to minimize the amount of nonsampling error introduced through nonresponse for sample housing units. One way of accomplishing this is by following up on mail nonrespondents during the CATI and CAPI phases.

MEASURES OF SAMPLING ERROR

Sampling error is the difference between an estimate based on a sample and the corresponding value that would be obtained if the estimate were based on the entire population (as from a census). Note that sample-based estimates will vary depending on the particular sample selected from the population. Measures of the magnitude of sampling error reflect the variation in the

estimates over all possible samples that could have been selected from the population using the same sampling methodology.

Estimates of the magnitude of sampling errors – in the form of margins of error – are provided with all published ACS data. The Census Bureau recommends that data users incorporate this information into their analyses, as sampling error in survey estimates could impact the conclusions drawn from the results.

Confidence Intervals and Margins of Error

Confidence Intervals – A sample estimate and its estimated standard error may be used to construct confidence intervals about the estimate. These intervals are ranges that will contain the average value of the estimated characteristic that results over all possible samples, with a known probability.

For example, if all possible samples that could result under the ACS sample design were independently selected and surveyed under the same conditions, and if the estimate and its estimated standard error were calculated for each of these samples, then:

1. Approximately 68 percent of the intervals from one estimated standard error below the estimate to one estimated standard error above the estimate would contain the average result from all possible samples;

2. Approximately 90 percent of the intervals from 1.645 times the estimated standard error below the estimate to 1.645 times the estimated standard error above the estimate would contain the average result from all possible samples.

3. Approximately 95 percent of the intervals from two estimated standard errors below the estimate to two estimated standard errors above the estimate would contain the average result from all possible samples.

The intervals are referred to as 68 percent, 90 percent, and 95 percent confidence intervals, respectively.

Margin of Error – Instead of providing the upper and lower confidence bounds in published ACS tables, the margin of error is provided instead. The margin of error is the difference between an estimate and its upper or lower confidence bound. Both the confidence bounds and the standard error can easily be computed from the margin of error. All ACS published margins of error are based on a 90 percent confidence level.

Standard Error = Margin of Error / 1.645

Lower Confidence Bound = Estimate - Margin of Error

Upper Confidence Bound = Estimate + Margin of Error

When constructing confidence bounds from the margin of error, the user should be aware of any "natural" limits on the bounds. For example, if a population estimate is near zero, the calculated value of the lower confidence bound may be negative. However, a negative number of people does not make sense, so the lower confidence bound should be reported as zero instead. However, for other estimates such as income, negative values do make sense. The context and meaning of the estimate must be kept in mind when creating these bounds. Another of these natural limits would be 100% for the upper bound of a percent estimate.

If the margin of error is displayed as '*****' (five asterisks), the estimate has been controlled to be equal to a fixed value and so has no sampling error. When using any of the formulas in the following section, use a standard error of zero for these controlled estimates.

Limitations –The user should be careful when computing and interpreting confidence intervals.

- The estimated standard errors (and thus margins of errors) included in these data products do not include portions of the variability due to nonsampling error that may be present in the data. In particular, the standard errors do not reflect the effect of correlated errors introduced by interviewers, coders, or other field or processing personnel. Nor do they reflect the error from imputed values due to missing responses. Thus, the standard errors calculated represent a lower bound of the total error. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence (i.e., 68, 90, or 95 percent). Thus, some care must be exercised in the interpretation of the data in this data product based on the estimated standard errors.

- Zero or small estimates; very large estimates — The value of almost all ACS characteristics is greater than or equal to zero by definition. For zero or small estimates, use of the method given previously for calculating confidence intervals relies on large sample theory, and may result in negative values which for most characteristics are not admissible. In this case the lower limit of the confidence interval is set to zero by default. A similar caution holds for estimates of totals close to a control total or estimated proportions near one, where the upper limit of the confidence interval is set to its largest admissible value. In these situations the level of confidence of the adjusted range of values is less than the prescribed confidence level.

CALCULATION OF STANDARD ERRORS

Direct estimates of the standard errors were calculated for all estimates reported in this product. The standard errors, in most cases, are calculated using a replicate-based methodology that takes into account the sample design and estimation procedures. Excluding the base weight, replicate weights were allowed to be negative in order to avoid underestimating the standard error. Exceptions include:

1. The estimate of the number or proportion of people, households, families, or housing units in a geographic area with a specific characteristic is zero. A special procedure is used to estimate the standard error.

2. There are no sample observations available to compute an estimate of a median, a proportion, or some other ratio, or an estimate of its standard error. The estimate is represented in the tables by "-" and the margin of error by "**" (two asterisks).

3. Only a small number of identical values are reported and used to calculate a median, aggregate, mean, or per capita amount. In this case, there are too few sample observations to compute a stable estimate of the standard error. The margin of error is represented in the tables by "**" (two asterisks).

4. The estimate of a median falls in the lower open-ended interval or upper open-ended interval of a distribution. If the median occurs in the lowest interval, then a "-" follows the estimate, and if the median occurs in the upper interval, then a "+" follows the estimate. In both cases the margin of error is represented in the tables by "***" (three asterisks).

Sums and Differences of Individual Estimates - The standard errors estimated from these tables are for individual estimates. Additional calculations are required to estimate the standard errors for sums of and differences between two sample estimates. The estimate of the standard error of a sum or difference is approximately the square root of the sum of the two individual standard errors squared; that is, for standard errors $SE(\hat{X})$ and $SE(\hat{Y})$ of estimates $\hat{X}$ and $\hat{Y}$:

$$SE(\hat{X} + \hat{Y}) = SE(\hat{X} - \hat{Y}) = \sqrt{[SE(\hat{X})]^2 + [SE(\hat{Y})]^2}$$

This method, however, will underestimate (overestimate) the standard error if the two items in a sum are highly positively (negatively) correlated or if the two items in a difference are highly negatively (positively) correlated.

Ratios – The statistic of interest may be the ratio of two estimates. First is the case where the numerator is not a subset of the denominator. The standard error of this ratio between two sample estimates is approximated as:

$$SE\left(\frac{\hat{X}}{\hat{Y}}\right) = \frac{1}{\hat{Y}}\sqrt{\left[SE(\hat{X})\right]^2 + \frac{\hat{X}^2}{\hat{Y}^2}\left[SE(\hat{Y})\right]^2}$$

Proportions/percents – For a proportion (or percent), a ratio where the numerator is a subset of the denominator, a slightly different estimator is used. Note the difference between the formulas for the standard error for proportions (below) and ratios (above) - the plus sign in the previous formula has been replaced with a minus sign. If the value under the square root sign is negative, use the ratio standard error formula above, instead. If $\hat{P} = \hat{X} / \hat{Y}$, then

If (P is the proportion and Q is its corresponding percent), then

$$SE(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[SE(\hat{X})]^2 - \frac{\hat{X}^2}{\hat{Y}^2}[SE(\hat{Y})]^2}$$

Percent Change – Calculating the percent change from one time period to another. For example, computing the percent change of a 2008-2010 estimate to a 2005-2007 estimate. Normally, the current estimate is compared to the older estimate.

Let the current estimate = X and the earlier estimate = Y, then the formula for percent change is:

$$SE\left(\frac{\hat{X} - \hat{Y}}{\hat{Y}}\right) = SE\left(\frac{\hat{X}}{\hat{Y}} - 1\right) = SE\left(\frac{\hat{X}}{\hat{Y}}\right)$$

This reduces to a ratio. The ratio formula above may be used to calculate the standard error.

Products – For a product of two estimates - for example if you want to estimate a proportion's numerator by multiplying the proportion by its denominator - the standard error can be approximated as

$$SE(\hat{X} \times \hat{Y}) = \sqrt{\hat{X}^2 \times [SE(\hat{Y})]^2 + \hat{Y}^2 \times [SE(\hat{X})]^2}$$

Comparing 1-Year Period Estimates with Overlapping 3-Year Period Estimates:

It should be noted that the 1-year and 3-year estimates represent period estimates. Due to the difficulty in interpreting the "difference" in period estimates of different lengths, the Census Bureau currently discourages users from making such comparisons.

$$SE(\hat{X}_{1-year} - \hat{Y}_{3-year}) = SE(\hat{Y}_{3-year} - \hat{X}_{1-year}) = \sqrt{\frac{1}{3}[SE(\hat{X}_{1-year})]^2 + [SE(\hat{Y}_{3-year})]^2}$$

TESTING FOR SIGNIFICANT DIFFERENCES

Users may conduct a statistical test to see if the difference between an ACS estimate and any other chosen estimates is statistically significant at a given confidence level. "Statistically significant" means that the difference is not likely due to random chance alone. With the two estimates (Est_1 and Est_2) and their respective standard errors (SE_1 and SE_2), calculate

$$Z = \frac{Est_1 - Est_2}{\sqrt{(SE_1)^2 + (SE_2)^2}}$$

If $Z < -1.645$ or $Z > 1.645$, then the difference between Est_1 and Est_2 can be said to be statistically significant at the 90% confidence level. Otherwise, the difference is not significant. This means that there is less than a 10 percent chance that the difference between these two estimates would be as large or larger by random chance alone.

Any estimate can be compared to an ACS estimate using this method, including other ACS estimates from the current year, the ACS estimate for the same characteristic and geographic area but from a previous year, Census 2000 100% counts and long form estimates, estimates from other Census Bureau surveys, and estimates from other sources. Not all estimates have sampling error — Census 2000 100% counts do not, for example, although Census 2000 long form estimates do — but they should be used if they exist to give the most accurate result of the test.

Users are also cautioned to not rely on looking at whether confidence intervals for two estimates overlap to determine statistical significance, because there are circumstances where that method will not give the correct test result. The Z calculation above is recommended in all cases.

EXAMPLES OF STANDARD ERROR CALCULATIONS

We will present some examples based on the real data to demonstrate the use of the formulas.

Example 1 - Calculating the Standard Error from the Confidence Interval

> The estimated number of males, never married is 39,124,581 from summary table B12001 for the period 2005-2007 in the United States. The margin of error is 73,194.
>
> Standard Error = Margin of Error / 1.645
>
> Calculating the standard error using the margin of error, we have:
>
> SE(39,124,581) = 73,194 / 1.645 = 44,495.

Example 2 - Calculating the Standard Error of a Sum

> We are interested in the number of people who have never married for the period 2005-2007 in the United States. From example 1, we know the number of males, never married is 39,124,581. From summary table B12001 we have the number of females, never married is 33,258,797 with a margin of error of 60,378. So, the estimated number of people who have never been married is 39,124,581 + 33,258,797 = 72,383,378. To calculate the standard error of this sum, we need the standard errors of the two estimates in the sum. We have the standard error for the number of males never married from example 1 as 44,495. The standard error for the number of females never married is calculated using the margin of error:
>
> SE(33,258,797) = 60,378 / 1.645 = 36,704.
>
> So using the formula for the standard error of a sum or difference we have:

$SE(72,383,378) = \sqrt{44,495^2 + 36,704^2} = 57,680.$

Caution: This method, however, will underestimate (overestimate) the standard error if the two items in a sum are highly positively (negatively) correlated or if the two items in a difference are highly negatively (positively) correlated.

To calculate the lower and upper bounds of the 90 percent confidence interval around 72,383,378 using the standard error, simply multiply 57,680 by 1.645, then add and subtract the product from 72,383,378. Thus the 90 percent confidence interval for this estimate is [72,383,378 - 1.645(57,680)] to [72,383,378 + 1.645(57,680)] or 72,288,494 to 72,478,262.

Example 3 - Calculating the Standard Error of a Percent

We are interested in the percentage of females who have never married to the number of people who have never married during the period of 2005-2007. The number of females, never married is 33,258,797 and the number of people who have never married is 72,383,378. To calculate the standard error of this sum, we need the standard errors of the two estimates in the sum. We have the standard error for the number of females never married from example 2 as 36,704 and the standard error for the number of people never married calculated from example 2 as 57,680.

The estimate is (33,258,797 / 72,383,378) * 100% = 45.95%

So, using the formula for the standard error of a proportion or percent, we have:

$$SE(45.95\%) = 100\% * \frac{1}{72,383,378}\sqrt{36,704^2 - \frac{33,258,797^2}{72,383,378^2} \times 57,680^2} = 0.04\%$$

To calculate the lower and upper bounds of the 90 percent confidence interval around 45.95 using the standard error, simply multiply 0.04 by 1.645, then add and subtract the product from 45.95. Thus the 90 percent confidence interval for this estimate is [45.95 - 1.645(0.04)] to [45.95 + 1.645(0.04)], or 45.89% to 46.01%.

Example 4 - Calculating the Standard Error of the Difference of Two Period Estimates

It should be noted that the 1-year and 3-year estimates represent period estimates. Due to the difficulty in interpreting the "difference" in period estimates of different lengths, the Census Bureau currently discourages users from making such comparisons.

We are interested in the "difference" of two estimates of the total population for age 3 and over in Wichita County, Texas. This can be found in table B14001 The estimated population for 2005 was 108,915 with a margin of error of 928. For 2005-2007, the

comparable estimate was 123,315 with a margin of error of 426, giving an estimated "difference" of 14,400.

To compute the standard error for the estimated "difference", we first compute the standard errors for the 2005 and 2005-2007 estimates by dividing the margins of error by 1.645, obtaining 564 and 259, respectively. The 2005-2007 data completely overlaps the 2005 estimate so we apply the formula,

$$SE(\hat{X}_{1-year} - \hat{Y}_{3-year}) = SE(\hat{Y}_{3-year} - \hat{X}_{1-year}) = \sqrt{\frac{1}{3}[SE(\hat{X}_{1-year})]^2 + [SE(\hat{Y}_{3-year})]^2}$$

$$= \sqrt{\frac{1}{3} * 564^2 + 259^2} = 416.$$

We get an estimated standard error for the "difference" of 416. To obtain a 90 percent confidence interval for the "difference", we multiply 416 by 1.645 to get 684, then add and subtract this result from the estimated difference of 14,400 to get a 90 percent confidence interval of (13,716, 15,084). Note that if we had ignored correcting to incorporate the correlation, the confidence interval would have been even wider.

CONTROL OF NONSAMPLING ERROR

As mentioned earlier, sample data are subject to nonsampling error. This component of error could introduce serious bias into the data, and the total error could increase dramatically over that which would result purely from sampling. While it is impossible to completely eliminate nonsampling error from a survey operation, the Census Bureau attempts to control the sources of such error during the collection and processing operations. Described below are the primary sources of nonsampling error and the programs instituted for control of this error. The success of these programs, however, is contingent upon how well the instructions were carried out during the survey.

- Coverage Error — It is possible for some sample housing units or persons to be missed entirely by the survey (undercoverage), but it is also possible for some sample housing units and persons to be counted more than once (overcoverage). Both the undercoverage and overcoverage of persons and housing units can introduce biases into the data, increase respondent burden and survey costs.

  A major way to avoid coverage error in a survey is to ensure that its sampling frame, for ACS an address list in each state, is as complete and accurate as possible. The source of addresses for the ACS is the MAF, which was created by combining the Delivery Sequence File of the United States Postal Service and the address list for Census 2000. An attempt is made to assign all appropriate geographic codes to each MAF address via an automated procedure using the Census Bureau TIGER (Topologically Integrated Geographic Encoding and Referencing) files. A manual coding operation based in the

appropriate regional offices is attempted for addresses, which could not be automatically coded. The MAF was used as the source of addresses for selecting sample housing units and mailing questionnaires. TIGER produced the location maps for CAPI assignments. Sometimes the MAF has an address that is the duplicate of another address already on the MAF. This could occur when there is a slight difference in the address such as 123 Main Street versus 123 Maine Street.

In the CATI and CAPI nonresponse follow-up phases, efforts were made to minimize the chances that housing units that were not part of the sample were interviewed in place of units in sample by mistake. If a CATI interviewer called a mail nonresponse case and was not able to reach the exact address, no interview was conducted and the case was eligible for CAPI. During CAPI follow-up, the interviewer had to locate the exact address for each sample housing unit. If the interviewer could not locate the exact sample unit in a multi-unit structure, or found a different number of units than expected, the interviewers were instructed to list the units in the building and follow a specific procedure to select a replacement sample unit. Person overcoverage can occur when an individual is included as a member of a housing unit but does not meet ACS residency rules.

Coverage rates give a measure of undercoverage or overcoverage of persons or housing units in a given geographic area. Rates below 100 percent indicate undercoverage, while rates above 100 percent indicate overcoverage. Coverage rates are released concurrent with the release of estimates on American FactFinder in the B98 series of detailed tables. Further information about ACS coverage rates may be found at http://www.census.gov/acs/www/UseData/sse/cov/cov_def.htm.

- Nonresponse Error — Survey nonresponse is a well-known source of nonsampling error. There are two types of nonresponse error – unit nonresponse and item nonresponse. Nonresponse errors affect survey estimates to varying levels depending on amount of nonresponse and the extent to which nonrespondents differ from respondents on the characteristics measured by the survey. The exact amount of nonresponse error or bias on an estimate is almost never known. Therefore, survey researchers generally rely on proxy measures, such as the nonresponse rate, to indicate the potential for nonresponse error.

    - Unit Nonresponse — Unit nonresponse is the failure to obtain data from housing units in the sample. Unit nonresponse may occur because households are unwilling or unable to participate, or because an interviewer is unable to make contact with a housing unit. Unit nonresponse is problematic when there are systematic or variable differences between interviewed and noninterviewed housing units on the characteristics measured by the survey. Nonresponse bias is introduced into an estimate when differences are systematic, while nonresponse error for an estimate evolves from variable differences between interviewed and noninterviewed households.

The ACS made every effort to minimize unit nonresponse, and thus, the potential for nonresponse error. First, the ACS used a combination of mail, CATI, and CAPI data collection modes to maximize response. The mail phase included a series of three to four mailings to encourage housing units to return the questionnaire. Subsequently, mail nonrespondents (for which phone numbers are available) were contacted by CATI for an interview. Finally, a subsample of the mail and telephone nonrespondents was contacted for by personal visit to attempt an interview. Combined, these three efforts resulted in a very high overall response rate for the ACS.

ACS response rates measure the percent of units with a completed interview. The higher the response rate, and consequently the lower the nonresponse rate, the less chance estimates may be affected by nonresponse bias. Response and nonresponse rates, as well as rates for specific types of nonresponse, are released concurrent with the release of estimates on American FactFinder in the B98 series of detailed tables. Further information about response and nonresponse rates may be found at http://www.census.gov/acs/www/UseData/sse/res/res_def.htm.

- Item Nonresponse — Nonresponse to particular questions on the survey questionnaire and instrument allows for the introduction of error or bias into the data, since the characteristics of the nonrespondents have not been observed and may differ from those reported by respondents. As a result, any imputation procedure using respondent data may not completely reflect this difference either at the elemental level (individual person or housing unit) or on average.

  Some protection against the introduction of large errors or biases is afforded by minimizing nonresponse. In the ACS, item nonresponse for the CATI and CAPI operations was minimized by the requirement that the automated instrument receive a response to each question before the next one could be asked. Questionnaires returned by mail were edited for completeness and acceptability. They were reviewed by computer for content omissions and population coverage. If necessary, a telephone follow-up was made to obtain missing information. Potential coverage errors were included in this follow-up.

  Allocation tables provide the weighted estimate of persons or housing units for which a value was imputed, as well as the total estimate of persons or housing units that were eligible to answer the question. The smaller the number of imputed responses, the lower the chance that the item nonresponse is contributing a bias to the estimates. Allocation tables are released concurrent with the release of estimates on American Factfinder in the B99 series of detailed tables with the overall allocation rates across all person and housing unit characteristics in the B98 series of detailed tables. Additional information on item nonresponse and allocations can be found at http://www.census.gov/acs/www/UseData/sse/ita/ita_def.htm

- Measurement and Processing Error — The person completing the questionnaire or responding to the questions posed by an interviewer could serve as a source of error,

although the questions were cognitively tested for phrasing, and detailed instructions for completing the questionnaire were provided to each household.

- Interviewer monitoring — The interviewer may misinterpret or otherwise incorrectly enter information given by a respondent; may fail to collect some of the information for a person or household; or may collect data for households that were not designated as part of the sample. To control these problems, the work of interviewers was monitored carefully. Field staff were prepared for their tasks by using specially developed training packages that included hands-on experience in using survey materials. A sample of the households interviewed by CAPI interviewers was reinterviewed to control for the possibility that interviewers may have fabricated data.

- Processing Error — The many phases involved in processing the survey data represent potential sources for the introduction of nonsampling error. The processing of the survey questionnaires includes the keying of data from completed questionnaires, automated clerical review, follow-up by telephone, manual coding of write-in responses, and automated data processing. The various field, coding and computer operations undergo a number of quality control checks to insure their accurate application.

- Content Editing — After data collection was completed, any remaining incomplete or inconsistent information was imputed during the final content edit of the collected data. Imputations, or computer assignments of acceptable codes in place of unacceptable entries or blanks, were needed most often when an entry for a given item was missing or when the information reported for a person or housing unit on that item was inconsistent with other information for that same person or housing unit. As in other surveys and previous censuses, the general procedure for changing unacceptable entries was to allocate an entry for a person or housing unit that was consistent with entries for persons or housing units with similar characteristics. Imputing acceptable values in place of blanks or unacceptable entries enhances the usefulness of the data.