# Assessment of Data Release Rules on the Reliability of Multiyear Estimates in American Community Survey Data Products

Michael D. Starsinic
U.S. Census Bureau, Washington, DC, 20233

## Abstract

The American Community Survey (ACS) uses data quality filtering to prevent the release of data products of sufficiently low quality. The ACS published its first 3-year period estimates in 2008 using sample data from 2005 through 2007. This paper examines the quality of the data products that are published for both the 2007 1-year period esimates and the 2005-2007 3-year period estimates, as well as the quantity of the data products that are not published. This paper also simulates filtering results and data quality characteristics for alternate methodologies that could improve the quality of published ACS data.

**Key Words:** American Community Survey, Estimate reliability, Filtering

## 1. Introduction[1]

The American Community Survey (ACS) is a continuous monthly survey that collects the data historically collected by the decennial census long form sample. Full implementation of the ACS began in January 2005, with the sample expanding to a size of approximately three million housing unit addresses, with sample in all counties and county equivalents in the 50 states, the District of Columbia, and Puerto Rico.

A single year's worth of sample in the ACS is not adequate to publish estimates for all geographic areas for which long form estimates were published in Census 2000. Instead, single-year estimates are published only for geographic areas with a population of at least 65,000. For smaller areas, several years of ACS sample are pooled together to create "period" estimates. The first estimates based on three years of pooled ACS data were published in 2008 for all areas with a population of at least 20,000 using data from 2005 through 2007. All geographic areas, including Census tracts and block groups, will be published using five years' worth of pooled ACS data. The five-year data will first be published in 2010 for the years 2005-2009. (U.S. Census Bureau 2009)

The ACS follows in the footsteps of the long form in publishing a very large array of data products accessible through the Census Bureau's American FactFinder (AFF) website. The ACS creates several thousand data products, some containing hundreds of individual estimates, for thousands of different geographic areas - over 6,000 areas for one-year data and over 13,000 for three-year data. That adds up to hundreds of millions of estimates released each year. The ACS realizes that not all the estimates that are produced are of

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

high quality - many may be based on a handful of sampled observations, and others are zero, with no sample cases in that geographic area having those characteristics.

The ACS has chosen to address this problem of low-quality data by instituting a process of "data quality filtering" for one-year and three-year data products, which identifies products with the highest concentrations of low-quality estimates and prevents their publication on AFF. This paper documents research that attempts to answer two questions about the ACS's filtering procedures:

- How does the current data quality filtering methodology affect the quality of the data that the ACS publishes, for both 1-year and 3-year data products?

- How do several alternate filtering methods affect the reliability of estimates that would be published under those rules?

## 2. ACS Data Products

Detailed tables are, as their name suggests, intended as the most finely detailed ACS data products, with tables crossing two, three, or four characteristics, such as age, sex, educational attainment, and income. Detailed tables are the building blocks of most other ACS data products. Estimates in data profiles, subject tables, and geographic comparison tables are all obtained either directly or indirectly (e.g. constructing percents from numerators and denominators) from detailed tables.

Certain detailed tables are published as "iterated" versions. In addition to the basic version of each of these tables, nine additional versions are created with additional race and/or Hispanic origin restrictions on the universe. The nine iterated groups are: white alone, black alone, Asian alone, Native Hawaiian and other Pacific islander alone, some other race alone, two or more races, Hispanic, and white alone not Hispanic. For example, a table on educational attainment might have the universe restricted to persons 25 years old and over. Iterated versions of this table would have universes of persons white alone 25 years old and over, black alone 25 years old and over, and so on.

Many, but not all, detailed tables also have "collapsed versions". These are modified from the "uncollapsed" version to reduce the number of lines (estimates) in the table. The purpose of defining collapsed versions was to create tables that may be more likely to pass the data quality filtering methodology should the uncollapsed version fail. The collapsed version may combine individual lines together – by combining separate age groups 15-17 and 18-21 into a single 15-21 group, for example – or may omit a dimension from the original table, such as age by sex by poverty being collapsed to age by poverty only. The layout of the collapsed table is prespecified and not done on-the-fly through an algorithm. Collapsed tables are also always produced, not just if the uncollapsed version fails the data quality filtering. Collapsed tables will be discussed further in the section 4.

## 3. Reliability, the ACS, and the Long Form

When assessing "data quality" and "reliability", the ACS generally looks at a measure of sampling error – the coefficient of variation (CV), which is defined as the standard error

of the estimate divided by the estimate itself. Estimates with smaller CVs are generally thought to have better data quality and higher reliability than estimates with larger CVs.

Reliability of published data is of heightened importance to the ACS compared to the long form for several reasons. Most importantly, all ACS estimates published on AFF are displayed with their 90 percent margin of error (the 90 percent level is the Census Bureau's standard), so a measure of an estimate's reliability is literally next to each estimate. In contrast, Census 2000 long form data are not displayed on AFF with any measure of sampling error. A user wishing to calculate an estimate's standard error must find the formula for the standard error in the SF3 Accuracy of the Data document, find the geographic area's percent-in-sample from AFF, obtain an appropriate design factor from published tables, and plug the data into the formula. (U.S. Census Bureau 2002) This is not an impossible task, but certainly not one a casual data user is likely to undertake. Estimates with high CVs could "hide" in the long form, unless the user calculated the standard error and the CV. Estimates with high CVs in the ACS are clearly visible to whomever looks for them.

The ACS is already at a disadvantage to the long form in the magnitude of sampling error due to the smaller sample size of the ACS. The ACS sample is fixed at about three million housing unit addresses each year, which is slightly more than two percent. Even the combined 5-year ACS sample of between 10 and 12 percent is smaller than the long form sample of about 16 percent.

Also, long form users knew that what they saw was all they would have for the next 10 years. If a user *needed* an estimate from the long form, the estimate's reliability was less of an issue because there were no other options. For the ACS, once the 5-year data begins publication in 2010, areas with populations above 65,000 will receive three estimates (1-year, 3-year, and 5-year periods) *each* year. Users looking at time series of ACS data (again, something unavailable to long form users) may see estimates bouncing around due to sampling error alone.

Most long form detailed tables from Census 2000's Summary File 3 (SF3) were made available down to the block group level. However, some tables, particularly those with many response categories (such as ancestry and language spoken at home) or those that were formed by crossing three or more characteristics, were only available down to the census tract level (one level higher than block group). These restrictions were made to address both confidentiality and data quality concerns.

## 4. Data Quality Filtering and Other Reliability Improvement Methods

The data quality filtering methodology used by the ACS is applied to each table separately for each geographic area eligible for publication. The filtering of a detailed table begins with the CV being calculated for each line in the table for a geographic area. If the median CV of all *detailed* lines in the table (those that are not the total line or a subtotal line) for the area is less than 0.61, then the detailed table passes the filtering process and will be published. If the median CV is greater than 0.61, then the table fails filtering (or, is filtered out), and will not be published on AFF for that geographic area. (U.S. Census Bureau 2009) The CV is undefined for a zero estimate, so for purposes of calculating the median CV, zero estimates are assigned a CV of one. This categorizes

zeroes as "poor quality" estimates, although that assumption is debatable, and we will address the issue later.

The cutoff value is set to 0.61 because, at that value (1/1.645 rounded to two decimal places), the 90 percent margin of error is equal to the estimate itself, and for larger CVs, the margin of error is larger than the estimate. In other words, for estimates with CVs of 0.61 or higher, the estimate is not significantly different from zero at the 90 percent confidence level. We are not attempting an actual statistical test here; clearly, if at least one sample respondent has a characteristic, the population count for that characteristic must be nonzero. This is simply a means of identifying – and giving a plausible statistical justification for – a reasonable cutoff value.

Detailed table filtering is applied at the table level, so either the whole table is published for a geographic area or the whole table is filtered out. Filtering at an estimate level would cause additional problems for users. Poor quality estimates are sprinkled throughout many otherwise-good tables. Further, complementary filtering might have to be applied to blank out estimates with acceptable CVs if the filtered estimates could be re-derived through subtraction. Filtering out these isolated cases would cause nightmares for any user attempting to add across geographic areas.

For other ACS data products that are built using data from detailed tables, such as data profiles and subject tables, filtering depends on the filtering of the underlying detailed table. If at least one table which feeds into a derived estimate is filtered out, then the derived estimate is filtered out as well.

Two other methods for controlling the release of low-reliability data have already been mentioned: population thresholds and collapsed tables. The one-year and three-year population publication thresholds of 65,000 and 20,000 have long been in place for ACS products (Alexander 1999). Some threshold needs to be in place – it would be unwise to try to publish data for all geographic areas based on just one year of ACS data. The exact thresholds were derived in the 1990s during early ACS testing and planning, using then-current assumptions about sample size and response rates. However, it's questionable whether those assumptions are still valid. If geographic areas near the thresholds have a higher proportion of poor quality estimates, then increasing the thresholds could be one way to improve the quality of the published data.

The collapsed detailed tables described in section 2 above are examples of the tension inherent in the filtering process. On one side is the Census Bureau, which wants to release the most reliable data. On the other side are the users who just want to see their data in tables that are being filtered out. The collapsing, by reducing the number of lines and increasing the number of cases in other lines, increases the likelihood that the collapsed table will pass filtering. The user can hopefully be satisfied with the reduced detail available in the collapsed table, if the alternative is having no data when the uncollapsed version fails filtering. We'll explore this tug-of-war later in this paper.

## 5. Analysis of Current Filtering Results

This analysis looks at over 1,150 detailed tables (including both collapsed and uncollapsed versions) that were created for *each* geographic area published in the 2007 1-year and 2005-2007 3-year ACS estimates. This includes most but not all detailed tables

published by the ACS. Some tables are not published with margins of error (such as allocation tables) and are not subject to filtering. Others are based on models and are not direct tabulations of ACS data. Still other excluded tables are based on something other than residence geography (place-of-work geography, for example).

## 5.1 Impact of Summary Levels Published

There were 6,566 geographic areas published for the 2007 1-year data, and 13,711 areas published for the 2005-2007 3-year data. Tables 1a and 1b show the distribution of selected geographic area types by population size range for the 1-year and 3-year period products.

Table 1a: Geographic Size Distribution for 1-Year 2007 Products

| | | Population Size Range | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Geo Type | Total | 65K 100K | 100K 125K | 125K 150K | 150K 200K | 200K 250K | 250K 500K | 500K 1M | > 1M |
| County | 800 | 228 | 91 | 76 | 103 | 54 | 124 | 86 | 38 |
| MCD | 187 | 99 | 37 | 17 | 11 | 6 | 12 | 4 | 1 |
| Place | 520 | 243 | 81 | 37 | 52 | 35 | 39 | 24 | 9 |
| School District | 950 | 405 | 145 | 87 | 104 | 64 | 95 | 34 | 16 |
| All Other Areas | 4,109 | 510 | 866 | 682 | 716 | 264 | 250 | 588 | 233 |
| Total | 6,566 | 1,485 | 1,220 | 899 | 986 | 423 | 520 | 736 | 297 |

Table 1b: Geographic Size Distribution for 3-Year 2005-2007 Products

| | | | | Population Size Range | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Geo Type | Total | < 65K | > 65K | 20K 25K | 25K 30K | 30K 35K | 35K 40K | 40K 45K | 45K 50K | 50K 55K | 55K 60K | 60K 65K |
| County | 1,882 | 1,087 | 795 | 247 | 180 | 128 | 145 | 121 | 99 | 57 | 56 | 54 |
| MCD | 999 | 812 | 187 | 228 | 163 | 114 | 78 | 67 | 52 | 44 | 41 | 25 |
| Place | 2,081 | 1,572 | 509 | 433 | 298 | 232 | 156 | 111 | 98 | 93 | 80 | 71 |
| School District | 3,298 | 2,380 | 918 | 654 | 467 | 305 | 264 | 211 | 145 | 128 | 113 | 93 |
| All Other Areas | 5,451 | 1,374 | 4,077 | 281 | 194 | 145 | 171 | 130 | 131 | 111 | 103 | 108 |
| Total | 13,711 | 7,225 | 6,486 | 1,843 | 1,302 | 924 | 814 | 640 | 525 | 433 | 393 | 351 |

MCD stands for “Minor Civil Division” and includes geographic entities such as townships. Place includes both incorporated places and unincorporated Census Designated Places. School district includes elementary, secondary, and unified districts. All Other Areas includes the nation, states, metropolitan and micropolitan areas, and Congressional Districts, among other types.

There was no attempt to unduplicate geographic areas that are represented in multiple categories. For example, the District of Columbia is included as a state, a county, a place, and a school district.

For both 1- and 3-year products, a sizable proportion of all published areas are in the smallest size categories. About 23 percent of 1-year areas are below 100,000, and about the same percentage of 3-year areas are below 30,000. Even very small changes in the population threshold would have a large impact on the number of eligible areas.

## 5.2 Impact on the Number of Tables and Estimates Published

Table 2a shows that for the 1-year data, just over five million tables were published, containing about 92 million estimates.

Table 2a: Filtering Characteristics of Tables and Estimates, 2007

| Pop Size Range (K) | Total Tables | Tables Published | Tables % Filtered | Total Estimates | Estimates Published | Estimates % Filtered |
|---|---|---|---|---|---|---|
| 65-100 | 1,649,817 | 969,366 | 41.2% | 37,321,362 | 15,918,606 | 57.3% |
| 100-125 | 1,355,403 | 861,858 | 36.4% | 30,661,363 | 14,856,500 | 51.5% |
| 125-150 | 998,775 | 659,060 | 34.0% | 22,593,934 | 11,588,084 | 48.7% |
| 150-200 | 1,095,433 | 757,372 | 30.9% | 24,780,399 | 13,698,710 | 44.7% |
| 200-250 | 469,951 | 340,401 | 27.6% | 10,630,874 | 6,385,795 | 39.9% |
| 250-500 | 577,712 | 447,332 | 22.6% | 13,068,792 | 8,880,953 | 32.0% |
| 500-1000 | 817,696 | 681,099 | 16.7% | 18,497,152 | 14,195,642 | 23.3% |
| 1000+ | 329,961 | 301,232 | 8.7% | 7,464,318 | 6,565,438 | 12.0% |
| Total | 7,294,748 | 5,017,720 | 31.2% | 165,018,194 | 92,089,728 | 44.2% |

About 31 percent of tables and about 44 percent of estimates were filtered out. For the areas with populations less than 100,000, 41 percent of tables and 57 percent of estimates were filtered. As one would expect, the percentages decline as the population size range increases. The difference between the percent of tables and estimates filtered out is due to tables with a larger number of estimates being more likely to be filtered out, especially for the smallest areas where we see the difference is the largest. Table 2b shows similar data for the filtering of 2005-2007 tables and estimates.

Table 2b: Filtering Characteristics of Tables and Estimates, 2005-2007

| Pop Size Range (K) | Total Tables | Tables Published | Tables % Filtered | Total Estimates | Estimates Published | Estimates % Filtered |
|---|---|---|---|---|---|---|
| 20-25 | 2,047,557 | 1,104,775 | 46.0% | 46,318,580 | 17,663,345 | 61.9% |
| 25-30 | 1,446,511 | 822,340 | 43.2% | 32,722,073 | 13,552,042 | 58.6% |
| 30-35 | 1,026,554 | 611,966 | 40.4% | 23,222,158 | 10,365,950 | 55.4% |
| 35-40 | 904,345 | 556,316 | 38.5% | 20,457,619 | 9,587,216 | 53.1% |
| 40-45 | 711,033 | 448,187 | 37.0% | 16,084,613 | 7,838,129 | 51.3% |
| 45-50 | 583,265 | 377,196 | 35.3% | 13,194,490 | 6,705,560 | 49.2% |
| 50-55 | 481,060 | 318,366 | 33.8% | 10,882,213 | 5,710,194 | 47.5% |
| 55-60 | 436,620 | 295,508 | 32.3% | 9,876,933 | 5,362,687 | 45.7% |
| 60-65 | 389,959 | 267,637 | 31.4% | 8,821,370 | 4,912,961 | 44.3% |
| 65-100 | 1,575,381 | 1,132,575 | 28.1% | 35,637,499 | 21,510,402 | 39.6% |
| 100-125 | 1,387,618 | 1,046,501 | 24.6% | 31,390,267 | 20,840,320 | 33.6% |
| 125-150 | 1,028,775 | 795,574 | 22.7% | 23,272,441 | 16,095,299 | 30.8% |
| 150-200 | 1,065,436 | 848,595 | 20.4% | 24,101,835 | 17,384,770 | 27.9% |
| 200-250 | 439,954 | 362,088 | 17.7% | 9,952,310 | 7,514,402 | 24.5% |
| 250-500 | 567,713 | 486,182 | 14.4% | 12,842,604 | 10,228,239 | 20.4% |
| 500-1000 | 814,363 | 730,054 | 10.4% | 18,421,756 | 15,717,024 | 14.7% |
| 1000+ | 326,628 | 309,359 | 5.3% | 7,388,922 | 6,860,125 | 7.2% |
| Total | 15,232,772 | 10,513,219 | 31.0% | 344,587,683 | 197,848,665 | 42.6% |

For the 3-year data, 10 million tables were published, containing about 198 million estimates. About 31 percent of tables and about 43 percent of estimates were filtered out – percentages very close to the 1-year data. For the areas with populations of less than 25,000, 46 percent of tables and 62 percent of estimates were filtered. The smallest size categories have somewhat higher filtering rates than the under 100,000 category for the 1-year data, but are in the same ballpark. When comparing the filtering rates for the areas above 65,000, about 1/3 fewer tables and estimates are filtered out for the 3-year data than the 1-year data across all size categories.

Table 3 shows that for many geographic areas, some (or most) of the nine race/Hispanic iteration groups may be very small. Therefore, it's not surprising that the iterated tables have much higher filtering rates than non-iterated tables.

Table 3: Filtering Characteristics of Iterated & Non-Iterated Tables and Estimates (for Selected Size Ranges)

| 1-Year | Pop Size Range (K) | Total Tables | Tables Published | Tables % Filtered | Total Estimates | Estimates Published | Estimates % Filtered |
|---|---|---|---|---|---|---|---|
| Not Iterated | 65-100 | 928,107 | 700,971 | 24.5% | 23,755,887 | 12,655,580 | 46.7% |
| | 1000+ | 185,619 | 181,033 | 2.5% | 4,751,223 | 4,584,819 | 3.5% |
| | Total | 4,103,672 | 3,474,728 | 15.3% | 105,037,784 | 71,698,187 | 31.7% |
| Iterated | 65-100 | 721,710 | 268,395 | 62.8% | 13,565,475 | 3,263,026 | 75.9% |
| | 1000+ | 144,342 | 120,199 | 16.7% | 2,713,095 | 1,980,619 | 27.0% |
| | Total | 3,191,076 | 1,542,992 | 51.6% | 59,980,410 | 20,391,541 | 66.0% |

| 3-Year | Pop Size Range (K) | Total Tables | Tables Published | Tables % Filtered | Total Estimates | Estimates Published | Estimates % Filtered |
|---|---|---|---|---|---|---|---|
| Not Iterated | 20-25 | 1,151,859 | 813,711 | 29.4% | 29,482,775 | 14,084,905 | 52.2% |
| | 1000+ | 183,744 | 180,350 | 1.8% | 4,703,232 | 4,608,165 | 2.0% |
| | Total | 8,569,226 | 7,264,145 | 15.2% | 219,337,698 | 153,941,817 | 29.8% |
| Iterated | 20-25 | 895,698 | 291,064 | 67.5% | 16,835,805 | 3,578,440 | 78.7% |
| | 1000+ | 142,884 | 129,009 | 9.7% | 2,685,690 | 2,251,960 | 16.1% |
| | Total | 6,663,546 | 3,249,074 | 51.2% | 125,249,985 | 43,906,848 | 64.9% |

More than half of all iterated tables are filtered out, while only about 15 percent of non-iterated tables are. A large majority of iterated tables and estimates are being filtered out for the smallest sized areas. Again, the overall filtering rates ("Total" lines) are very close when comparing 1-year against 3-year.

## 5.3 Impact on Filtering on the Reliability of Published Estimates

Table 4a gives the distribution of the size of the CV for all 92 million *published* estimates for the 1-year data.

Table 4a: CV Distribution of Published Estimates, 2007

| Pop Size Range (K) | Tot Est | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0* |
|---|---|---|---|---|---|---|---|---|---|
| 65-100 | 15,918,606 | 22.4% | 19.9% | 15.2% | 10.7% | 7.4% | 5.7% | 11.1% | 7.5% |
| 100-125 | 14,856,500 | 24.4% | 21.3% | 15.2% | 10.1% | 6.8% | 5.3% | 10.3% | 6.6% |
| 125-150 | 11,588,084 | 25.9% | 21.6% | 14.9% | 9.8% | 6.5% | 5.1% | 9.9% | 6.3% |
| 150-200 | 13,698,710 | 27.7% | 22.0% | 14.6% | 9.3% | 6.2% | 4.9% | 9.5% | 5.8% |
| 200-250 | 6,385,795 | 29.1% | 22.2% | 14.2% | 9.0% | 6.0% | 4.7% | 9.1% | 5.6% |
| 250-500 | 8,880,953 | 33.6% | 22.2% | 13.0% | 8.2% | 5.5% | 4.3% | 8.1% | 5.0% |
| 500-1000 | 14,195,642 | 40.8% | 21.9% | 11.6% | 7.2% | 4.7% | 3.5% | 6.4% | 3.9% |
| 1000+ | 6,565,438 | 57.8% | 17.6% | 8.5% | 4.8% | 2.9% | 2.1% | 3.9% | 2.4% |
| Total | 92,089,728 | 30.9% | 21.2% | 13.8% | 9.0% | 6.0% | 4.6% | 8.9% | 5.7% |

The "est=0*" category in Table 4a primarily contains zero estimates, but also includes certain special cases for median and ratio estimates where either the estimate or the standard error could not be calculated (hence the asterisk). As the population ranges increase, the proportion of estimates with small CVs increases. About 22 percent of the estimates in the 65,000 to 100,000 size category have CVs less than 0.1, and about 42 percent in that category have CVs less than 0.2. In the same size category, about 19 percent of published estimates either have a CV greater than 0.61 or is a zero estimate. Even in the largest areas, there are still more than six percent of published estimates that are zero or have a CV greater than 0.61. Over the 92 million total estimates, 52 percent have CVs less than 0.2, nine percent have CVs greater than 0.61, and six percent are zero.

Table 4b: CV Distribution of Published Estimates, 2005-2007

| Pop Size Range (K) | Tot Est | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0* |
|---|---|---|---|---|---|---|---|---|---|
| 20-25 | 17,663,345 | 22.9% | 18.9% | 14.7% | 10.5% | 7.3% | 5.7% | 11.5% | 8.3% |
| 25-30 | 13,552,042 | 24.0% | 19.5% | 14.7% | 10.3% | 7.1% | 5.5% | 11.1% | 7.8% |
| 30-35 | 10,365,950 | 24.8% | 20.2% | 14.7% | 10.1% | 6.9% | 5.3% | 10.6% | 7.3% |
| 35-40 | 9,587,216 | 26.2% | 20.6% | 14.6% | 9.8% | 6.6% | 5.1% | 10.3% | 6.8% |
| 40-45 | 7,838,129 | 26.9% | 20.9% | 14.6% | 9.6% | 6.5% | 5.0% | 10.0% | 6.6% |
| 45-50 | 6,705,560 | 28.0% | 21.1% | 14.4% | 9.4% | 6.2% | 4.9% | 9.7% | 6.3% |
| 50-55 | 5,710,194 | 27.9% | 21.6% | 14.4% | 9.3% | 6.2% | 4.8% | 9.5% | 6.2% |
| 55-60 | 5,362,687 | 28.3% | 21.8% | 14.4% | 9.2% | 6.1% | 4.8% | 9.4% | 6.0% |
| 60-65 | 4,912,961 | 29.2% | 21.9% | 14.1% | 9.0% | 6.0% | 4.7% | 9.2% | 5.9% |
| 65-100 | 21,510,402 | 30.9% | 22.3% | 13.7% | 8.6% | 5.8% | 4.5% | 8.8% | 5.6% |
| 100-125 | 20,840,320 | 33.2% | 22.6% | 13.0% | 8.1% | 5.5% | 4.3% | 8.1% | 5.2% |
| 125-150 | 16,095,299 | 34.9% | 22.5% | 12.7% | 7.9% | 5.3% | 4.1% | 7.8% | 4.9% |
| 150-200 | 17,384,770 | 37.2% | 22.3% | 12.2% | 7.6% | 5.1% | 3.9% | 7.3% | 4.6% |
| 200-250 | 7,514,402 | 39.4% | 22.0% | 11.7% | 7.4% | 4.9% | 3.6% | 6.7% | 4.2% |
| 250-500 | 10,228,239 | 45.1% | 20.9% | 10.9% | 6.7% | 4.2% | 3.0% | 5.6% | 3.5% |
| 500-1000 | 15,717,024 | 53.0% | 19.1% | 9.6% | 5.5% | 3.3% | 2.4% | 4.5% | 2.7% |
| 1000+ | 6,860,125 | 68.4% | 14.5% | 6.2% | 3.3% | 2.0% | 1.4% | 2.7% | 1.6% |
| < 65 | 81,698,084 | 25.6% | 20.3% | 14.6% | 9.9% | 6.8% | 5.2% | 10.5% | 7.1% |
| 65+ | 116,150,581 | 39.8% | 21.3% | 11.8% | 7.3% | 4.8% | 3.6% | 6.9% | 4.4% |
| Total | 197,848,665 | 33.9% | 20.9% | 13.0% | 8.4% | 5.6% | 4.3% | 8.4% | 5.5% |

Overall, the CV distribution of published estimates is slightly better for the 3-year data than for the 1-year data, as seen in Table 4b. The percent of estimates with a CV less than 0.1 and less than 0.2 are both about 3 percentage points higher that is seen for the 1-year data. The percent of estimates with a CV greater than 0.61 is also about half a percentage point lower. The distribution for the areas with populations of less than 25,000 is very similar to the distribution for areas with populations of less than 100,000 for the 1-year data.

For areas with populations of less than 65,000, overall about 45 percent of areas have CVs less than 0.2, while a little less than 18 percent have a CV greater than 0.61 or are equal to zero. Comparing the 1-year and 3-year distributions for areas with populations greater than 65,000, the percent of estimates with a CV less than 0.1 and less than 0.2 both increased by about 9 percentage points, but the percentage point decreases in poor quality estimates were not very sizable. Although the CVs for all estimates in the tables that were published for the 1-year data should improve in the 3-year tabulations, some tables that failed filtering under 1-year are now passing under 3-year, adding a new batch of high-CV and zero estimates.

It is interesting to note that, even for areas with a population greater than one million, 2.7 percent of *published* estimates *still* had CVs above 0.61, and another 1.6 percent were zeroes.

It is also important to ask whether the current rules end up filtering out a lot of high quality estimates.

Table 5: CV Distribution for Filtered-Out Estimates (for Selected Size Ranges)

| Period | Pop Size Range (K) | Tot Est | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1-Year | 65-100 | 21,402,756 | 3.0% | 4.6% | 5.1% | 4.9% | 4.7% | 5.0% | 22.2% | 50.5% |
| | 1000+ | 898,880 | 3.6% | 4.8% | 5.0% | 5.2% | 5.4% | 6.0% | 26.0% | 44.0% |
| | Total | 72,928,466 | 3.0% | 4.9% | 5.2% | 5.1% | 4.8% | 5.2% | 23.1% | 48.7% |
| 3-Year | 20-25 | 28,655,235 | 3.4% | 4.3% | 4.6% | 4.5% | 4.3% | 4.5% | 20.8% | 53.5% |
| | 1000+ | 528,797 | 4.5% | 5.0% | 4.9% | 5.4% | 5.7% | 6.1% | 25.9% | 42.4% |
| | Total | 146,739,018 | 3.2% | 4.6% | 4.9% | 4.8% | 4.6% | 5.0% | 22.5% | 50.3% |

Table 5 shows that very little high quality data is removed. The filtered-out data consists of about half zero estimates, about another quarter of estimates with very high CVs, and only three percent have a CV less than 0.1.

For completeness, Table 6 shows CV distributions for selected ranges with no filtering applied. With no filtering, more published estimates would be zero or have a CV greater than 0.61 than would have a CV less than 0.2.

Table 6: CV Distribution for Published and Filtered-Out Estimates Combined (for Selected Size Ranges)

| Period | Range | Tot Est | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1-Year | 65-100 | 37,321,362 | 11.3% | 11.1% | 9.4% | 7.4% | 5.8% | 5.3% | 17.5% | 32.1% |
| | 1000+ | 7,464,318 | 51.3% | 16.0% | 8.1% | 4.8% | 3.2% | 2.6% | 6.5% | 7.4% |
| | Total | 165,018,194 | 18.6% | 14.0% | 10.0% | 7.3% | 5.5% | 4.9% | 15.2% | 24.7% |
| 3-Year | 20-25 | 46,318,580 | 10.8% | 9.9% | 8.5% | 6.8% | 5.4% | 5.0% | 17.3% | 36.3% |
| | 1000+ | 7,388,922 | 63.8% | 13.8% | 6.1% | 3.4% | 2.3% | 1.7% | 4.3% | 4.5% |
| | Total | 344,587,683 | 20.9% | 14.0% | 9.6% | 6.9% | 5.2% | 4.6% | 14.4% | 24.6% |

## 6. Analysis of Alternate Filtering Rules

There are several ways to adjust the number of tables and estimates that would be published. One way is to adjust the filtering rules, and another is to adjust the publication thresholds. This section will simulate possible outcomes for new rules by applying them to the available 2007 1-year and 2005-2007 3-year detailed tables. For these simulations, only tables that are counts of persons, households, families, and housing units will be used. CV behavior is well understood for estimates like these, but somewhat less so for the other types of detailed table estimates: ratios, aggregates (totals of quantities other than persons, households, or housing units, such as total travel time to work or total household income), and medians. All estimates in a single detailed table are of the same type, and no detailed table contains proportions.

### 6.1 Alternative Table Filtering Rules

We will consider the following modified filtering rules, in addition to the current methodology (which will be noted as CUR in Tables 7a and 7b):

1. No filtering. (NONE)

2. Filter iterated tables only (using the current methodology), but do not filter non-iterated tables. As we saw in Table 3, iterated tables are filtered at a much higher rate than non-iterated tables. (ITER)

3. Change the rule from "median CV > 0.61" to "Q1 CV > 0.61", where Q1 is the first quartile or the 25th percentile. The current rule basically requires half or more estimates to have CVs greater than 0.61 or be zero for the table to be filtered. This rule would instead require roughly three-quarters or more of the estimates to have CVs greater than 0.61 or be zero for the table to be filtered. (Q1)

4. Ignore zeroes, instead of assigning a CV of one. For example, if there are 10 estimates in a table, of which three are zero, then the median CV would be calculated on the seven non-zero estimates only. Zero is a perfectly legitimate result – no one in the sample had the designated characteristic, and they may not bother users as much as a non-zero estimate with a high CV. (ZERO)

5. Change the rule from "median CV > 0.61" to "median CV > 0.50". (0.5)

6. Change the rule from "median CV > 0.61" to "median CV > 0.40". Rules 5 & 6 are stricter than the current methodology, unlike rules 1 through 4, and will filter out more estimates and tables. We want to see how much of an effect this has on the number of tables and estimates published, as well as improving the CV distribution. (0.4)

Table 7a shows the filtering rates and CV distribution for the current method and the six alternative methods.

Table 7a: Filtering Rates and CV Distribution for Alternate Filtering Rules (Count Estimates Only), 2007

| Rule | Tables % Filtered | Estimates % Filtered | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0 |
|---|---|---|---|---|---|---|---|---|---|---|
| CUR | 36.7% | 46.2% | 29.7% | 21.2% | 14.0% | 9.2% | 6.2% | 4.8% | 9.3% | 5.5% |
| NONE | 0.0% | 0.0% | 18.6% | 14.0% | 10.0% | 7.3% | 5.5% | 4.9% | 15.2% | 24.7% |
| ITER | 22.6% | 24.0% | 24.1% | 17.7% | 12.1% | 8.4% | 6.0% | 5.0% | 12.8% | 13.9% |
| Q1 | 24.0% | 27.7% | 23.7% | 18.3% | 12.9% | 9.1% | 6.6% | 5.5% | 13.1% | 10.7% |
| ZERO | 26.5% | 30.0% | 24.4% | 18.6% | 12.9% | 9.0% | 6.4% | 5.3% | 11.8% | 11.7% |
| 0.5 | 42.7% | 53.2% | 32.5% | 22.3% | 14.3% | 9.0% | 5.8% | 4.0% | 7.7% | 4.4% |
| 0.4 | 49.7% | 61.3% | 36.3% | 23.6% | 14.3% | 8.4% | 4.7% | 3.3% | 6.1% | 3.3% |

One surprising result from the alternate filtering rules seen in Table 7a is how similar rules 2 through 4 are in both filtering rates and CV distribution. The rationales are different, but the overall results are very close. All three have table filtering rates around 25 percent, and "low quality" rates of about 25 percent as well. Any of the three would publish more tables and estimates, at the cost of decreasing the current overall reliability.

Rules 5 and 6, which lowered the CV cutoff threshold, did improve the CV distribution, but at a fairly steep cost in filtering rates, with rule 6 filtering out almost half of all tables.

Table 7b: Filtering Rates and CV Distribution for Alternate Filtering Rules (Count Estimates Only), 2005-2007

| Rule | Tables % Filtered | Estimates % Filtered | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0 |
|---|---|---|---|---|---|---|---|---|---|---|
| CUR | 36.4% | 44.5% | 32.8% | 21.0% | 13.3% | 8.6% | 5.8% | 4.5% | 8.7% | 5.4% |
| NONE | 0.0% | 0.0% | 20.9% | 14.0% | 9.6% | 6.9% | 5.2% | 4.6% | 14.4% | 24.6% |
| ITER | 22.4% | 23.6% | 27.0% | 17.6% | 11.5% | 7.9% | 5.6% | 4.7% | 12.0% | 13.6% |
| Q1 | 24.3% | 27.6% | 26.7% | 18.4% | 12.4% | 8.6% | 6.2% | 5.2% | 12.3% | 10.2% |
| ZERO | 26.5% | 29.2% | 27.3% | 18.5% | 12.3% | 8.5% | 6.0% | 4.9% | 11.1% | 11.4% |
| 0.5 | 41.9% | 51.2% | 35.7% | 22.0% | 13.4% | 8.4% | 5.4% | 3.7% | 7.2% | 4.2% |
| 0.4 | 48.3% | 58.8% | 39.6% | 23.2% | 13.3% | 7.8% | 4.3% | 3.0% | 5.6% | 3.2% |

The 3-year results in Table 7b are very similar to the 1-year results in Table 7a, and the same conclusions can be drawn.

## 6.2 Alternative Publication Threshold Restrictions

The second way to adjust the CV distribution is to increase the population threshold values from the current values of 65,000 and 20,000. As we saw in Tables 4a and 4b, estimates from the smaller population size groups had overall lower reliability than the larger groups. By raising the threshold, tables and estimates from below that threshold are now "filtered out", and the overall reliability profile for the remaining published data would improve. Table 8a applies this method to the 1-year data. The first row is the current threshold of 65,000, and the threshold increases up to 1,000,000 as you go down the table.

Table 8a: CV Distribution for Alternate Population Threshold Values (Count Estimates Only), 2007

| Threshold | Tot Est | # Geo | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 65K+ | 84,601,455 | 6,566 | 29.7% | 21.2% | 14.0% | 9.2% | 6.2% | 4.8% | 9.3% | 5.5% |
| 100K+ | 70,337,507 | 5,081 | 31.4% | 21.5% | 13.7% | 8.9% | 5.9% | 4.6% | 8.8% | 5.2% |
| 125K+ | 56,865,036 | 3,861 | 33.4% | 21.5% | 13.3% | 8.5% | 5.7% | 4.4% | 8.3% | 5.0% |
| 150K+ | 46,301,182 | 2,962 | 35.4% | 21.5% | 12.9% | 8.1% | 5.4% | 4.2% | 7.9% | 4.7% |
| 200K+ | 33,732,421 | 1,976 | 38.7% | 21.4% | 12.1% | 7.6% | 5.0% | 3.8% | 7.1% | 4.3% |
| 250K+ | 27,833,795 | 1,553 | 41.0% | 21.2% | 11.6% | 7.2% | 4.7% | 3.6% | 6.6% | 4.0% |
| 500K+ | 19,555,727 | 1,033 | 44.7% | 20.8% | 10.9% | 6.7% | 4.3% | 3.2% | 5.8% | 3.6% |
| 1,000K+ | 6,217,705 | 297 | 56.4% | 18.0% | 8.8% | 5.0% | 3.0% | 2.2% | 4.0% | 2.5% |

Comparing Tables 7a and 8a, we can see that filtering rule 5 has a CV distribution between the 100,000 and 125,000 thresholds. Likewise, rule 6's CV distribution is between the 150,000 and 200,000 thresholds.

One new cost of this method is the reduction in the number of geographic areas that would receive 1-year and 3-year data. By raising the 1-year threshold from 65,000 to 100,000, about 1,500 geographic areas would cease receiving 1-year estimates.

Note again that even with an unreasonably high threshold of one million, there are a fair number of high CV and zero estimates still being published.

Similar results for 3-year threshold values can be seen in Table 8b.

Table 8b: CV Distribution for Alternate Population Threshold Values (Count Estimates Only), 2005-2007

| Threshold | Tot Est | # Geo | cv<.1 | .1<cv<.2 | .2<cv<.3 | .3<cv<.4 | .4<cv<.5 | .5<cv<.61 | cv>.61 | est=0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 20K+ | 182,226,303 | 13,711 | 32.8% | 21.0% | 13.3% | 8.6% | 5.8% | 4.5% | 8.7% | 5.4% |
| 25K+ | 166,587,975 | 11,868 | 33.8% | 21.2% | 13.1% | 8.4% | 5.6% | 4.3% | 8.4% | 5.2% |
| 30K+ | 154,483,623 | 10,566 | 34.7% | 21.3% | 12.9% | 8.2% | 5.5% | 4.2% | 8.1% | 5.0% |
| 35K+ | 145,153,216 | 9,642 | 35.4% | 21.4% | 12.8% | 8.1% | 5.4% | 4.1% | 8.0% | 4.9% |
| 40K+ | 136,484,925 | 8,828 | 36.0% | 21.5% | 12.7% | 7.9% | 5.3% | 4.1% | 7.8% | 4.8% |
| 45K+ | 129,371,961 | 8,188 | 36.6% | 21.5% | 12.5% | 7.8% | 5.2% | 4.0% | 7.6% | 4.7% |
| 50K+ | 123,263,346 | 7,663 | 37.1% | 21.5% | 12.4% | 7.7% | 5.1% | 3.9% | 7.5% | 4.6% |
| 55K+ | 118,047,549 | 7,230 | 37.5% | 21.5% | 12.3% | 7.7% | 5.1% | 3.9% | 7.4% | 4.6% |
| 60K+ | 113,134,226 | 6,837 | 38.0% | 21.5% | 12.2% | 7.6% | 5.0% | 3.8% | 7.3% | 4.5% |
| 65K+ | 108,623,152 | 6,486 | 38.4% | 21.5% | 12.1% | 7.5% | 5.0% | 3.8% | 7.2% | 4.4% |
| 100K+ | 88,745,791 | 5,068 | 40.4% | 21.3% | 11.7% | 7.2% | 4.8% | 3.6% | 6.8% | 4.2% |
| 125K+ | 69,350,882 | 3,819 | 42.8% | 20.9% | 11.3% | 6.9% | 4.5% | 3.4% | 6.3% | 3.9% |
| 150K+ | 54,329,231 | 2,893 | 45.4% | 20.5% | 10.8% | 6.6% | 4.2% | 3.1% | 5.8% | 3.6% |
| 200K+ | 38,059,986 | 1,934 | 49.6% | 19.7% | 10.1% | 6.0% | 3.8% | 2.7% | 5.1% | 3.1% |
| 250K+ | 31,007,272 | 1,538 | 52.3% | 19.1% | 9.6% | 5.6% | 3.5% | 2.5% | 4.6% | 2.8% |
| 500K+ | 21,375,354 | 1,027 | 56.2% | 18.1% | 8.9% | 5.0% | 3.0% | 2.2% | 4.1% | 2.5% |
| 1,000K+ | 6,515,499 | 294 | 67.2% | 14.9% | 6.4% | 3.4% | 2.1% | 1.5% | 2.8% | 1.7% |

## 7. Conclusions

Current filtering rules allow publication of about 92 million estimates for the 1-year data and about 198 million for the 3-year data, but still filter out about 31 percent of tables and 43 percent of estimates (for both periods).

The CV distributions for both periods are also quite similar, with a little more than half of published estimates having CVs less than 0.2, and another 14 percent having CVs greater than 0.61 or being zeroes. Applying no data quality filtering would dramatically increase the amount of poor-quality data that would be published. Also, the data that is not published due to current filtering rules contains very few low-CV estimates.

Both altering the filtering rules to make them more restrictive and raising the data publication threshold allow for an improvement in quality for published data. A major drawback to both options is taking away estimates from users (either by filtering out tables that were published, or not publishing any estimates for their geographic area), who would likely be vocal in their complaints.

The goal of this research was *not* to make a recommendation on specific modifications to the filtering rules or the publication thresholds to achieve a certain goal. It was instead intended to document the effects of the current rules and provide simulated data based on various alternatives. There are many measures that can be chosen to rank possible new rules, but the choice of which to use is largely subjective. It would be up to the decision-makers for the ACS how to use the data presented here to make any decisions on changing the existing rules.

## References

Alexander, Charles H. (1998), "Recent developments in the American Community Survey", ASA Proceedings of the Section on Survey Research Methods, 92-100, American Statistical Association (Alexandria, VA)

U.S. Census Bureau (2002a), "Summary File 3 Technical Documentation", http://www.census.gov/prod/cen2000/doc/sf3.pdf.

U.S. Census Bureau (2009), "Design and Methodology: American Community Survey", http://www.census.gov/acs/www/Downloads/dm1.pdf