Errata for CTPP Part 1

by Phil Salopek, Census Bureau, November 17, 2003

During the review of the initial release of Part 1 by the State Departments of Transportation and the Metropolitan Planning Organizations, as well as our own internal review, the following errors were identified. These errors will be corrected before distribution of the final Part 1 files begins in January.

Errors were identified in the following tables. Most of the problems are minor except for tables 39, 80, and 106, where significant errors were found. Unless otherwise stated, the errors occurred in both the ASCII data files and in the data bundled with the CTPP Access Tool (CAT) software.

1) Table 33 - the wrong universe was tabulated. The table will be re-run for all workers. The correction should not result in large changes.

2) Table 36 - the wrong universe was tabulated. The table will be re-run for workers for whom poverty status has been determined. The correction should not result in large changes.

3) Table 39 - the data were read into the software incorrectly, and mislabeled as means of transportation instead of worker earnings. This error did not occur in the ASCII files.

4) Table 40 - same as table 36.

5) Table 46 - same as table 36.

6) Table 80 - an error in the record layout for the ASCII data switched the second and third dimensions of the table and caused it to be read into the software incorrectly. Users of the ASCII file were given the incorrect record layout too and probably read the table incorrectly as well. The error is obvious in the Race/Hispanic data if it was read-in wrong.

7) Table 94 - the wrong universe was tabulated. This affects primarily summary level 930 (MPO region total), but we advise you to spot-check other summary levels for changes as well when you receive the final data. The table will be re-run for workers with earnings in 1999. The correction should not result in large changes to the median earnings data.

8) Table 95 - the wrong universe was tabulated. This affects summary level 930, but may also affect the block group (150) and taz (940) summary levels. We advise you to check the other summary levels when you receive the final data. The table will be re-run for workers with earnings in 1999 residing in households. The correction should not result in large changes to the median earnings data.

9) Table 106 - the array indices were incorrectly specified in the tabulation of this table and as a result the data are all wrong. Do not use this table from the initial release of Part 1.

In addition to the errors in specific tables noted above, there were some additional, more general errors in the initial release of Part 1.

10) For all tables in four summary levels: MSA/CMSA (380), CMSA-PMSA (385), Urban area (400), and MPO region (930) there was an error in the data for some of the geographic areas that crossed state boundaries. The error consisted of not including the input files for all the states that made up the area in the data tables. For example, if an MPO consisted of three counties, each in a different state, then the intention was that the MPO region summary level would provide data for the entire MPO, that is, the sum of the three counties, using the input file from each state. This was sometimes done correctly, but not always. However, there is no consistent pattern to the error. One way to check the four summary levels is to compare them to the sum of the individual county-level data for the counties making up the multi-state area.

11) During our creation of the ASCII files, as a result of merging person tables, household tables, and housing unit tables into the same data file some extraneous records were created. The software vendors deleted the records so we don't think the data from the software CDs will be affected. However, to clean up the ASCII files we are re-running tables 47 through 87 for all states to make sure no extraneous records are included in the final release.

12) For some states there were no PUMA boundaries or shape files included on the software CDs. This will be corrected so that all PUMA boundaries are provided.

13) Some of the MSA/CMSA and PMSA names shown in the initial release of Part 1 were incorrect. These will be corrected in the final release.

14) In the first few states released, the mean and standard deviation of travel time were incorrectly calculated by including workers who worked at home. The errata file for each of these states documented the error and a work-around. This will be correct for all states in the final release.

15) For most of the single-cell tables in the initial release of Part 1 the column headers in the CTPP browser were incorrect. For example, the label may have said all persons when the table in fact was a tally of all housing units. This will be corrected in the final release.

The following notes about the data in Part 1 may also be helpful to data users.

i) Rounding. All data in the CTPP have been rounded according to the following procedure:

- Estimates of 0 remain 0

- Estimated values from 1 through 7 are rounded to 4
- Estimated values of 8 or greater are rounded to the nearest multiple of 5

Totals are rounded independently of the detailed cells making up the total, so totals may not equal the sum of the categories. The difference between the total and the sum of the individual categories can be substantial for tables with more than 30 cells.

ii) CTPP 2000 is based on the long-form (sample) data from Census 2000. The sample observations are weighted or inflated to represent the total population. Population controls are used in the weighting process at the "weighting area" level. Geographic units (cities, tracts, etc.) smaller than a weighting area can show large differences in total population from the Summary File 1 (SF1) figures. See the discussion in the Summary File 3 Technical Documentation, Chapter 9, Data Notes 6 and 7 (attached). Therefore, CTPP data may not agree with the hundred percent (short-form) data from SF1 or the PL 94-171 (redistricting) files. Differences for small geographic units may be particularly large. The best comparison for CTPP estimates is to look at SF3 numbers for the same geographic unit, although the CTPP numbers will of course be rounded.

iii) The count of households in Census 2000 and CTPP is not the same as the count of occupied housing units. There are two weights used in the Census weighting process, a housing unit weight and a person weight. Data for households from the census and CTPP are not derived using the housing unit weight, but rather, use the person weight of the householder. This is particularly important to note when a variable is tabulated for one universe in the census and tabulated for a different universe in CTPP. A good example of this is vehicles available. In the census it is always tabulated for housing units, but in the CTPP we tabulate it for households. Thus the CTPP number will look different than the Census 2000 number.

iv) Not all geographic areas have data available for them in CTPP 2000. Some areas have zero housing units and zero population living in them and so will not show up in Part 1 of CTPP. However, some of these areas may have people working in them and so will show up in Part 2. In addition, since the long form data are collected on a sample basis, it is possible that none of the people living or working in an area were picked up in the census sample. In these cases there will again not be any data for the area provided in the CTPP. Note, however, that the boundary files used in the CAT software are complete. So all the areas will show up on the maps, in both the geographic selection tool and in the mapping tool, but in some cases there will be no data to display in the data browser or on the map.

v) Not all summary levels are available for all geographic areas. For example, the TAZ summary level is only available in counties where TAZs were defined. Similarly, block group data are only available for counties where the State DOT

or MPO specifically requested that level of geography. For TAZs, it is almost always the case that if there are TAZ boundaries shown in the CAT software geographic selection tool, then data for TAZs is present. However, there may be a small number of instances where TAZs were defined and the boundaries show up on the maps in the CAT software, but there is no data available because the MPO decided to request data for a different geographic level. For block groups the situation is much different. All the block group boundaries in TIGER were imported into the CAT software, regardless of whether data were going to be produced for them or not. As a result, the block group summary level always appears in the geographic selection tool, and there are always boundaries shown for them on the map. However, for most areas across the country there will not be data available at the block group level. Currently there is no warning about the absence of data for a particular summary level, nor any way for the user to tell if data are available. We are working with the software vendors to resolve this issue in future releases.

Summary File 3
Data Note 6

COMPARING SF 3 ESTIMATES WITH CORRESPONDING VALUES IN SF 1 AND SF 2

As in earlier censuses, the responses from the sample of households reporting on long forms must be weighted to reflect the entire population. Specifically, each responding household represents, on average, six or seven other households who reported using short forms.

One consequence of the weighting procedures is that each estimate based on the long form responses has an associated confidence interval. These confidence intervals are wider (as a percentage of the estimate) for geographic areas with smaller populations and for characteristics that occur less frequently in the area being examined (such as the proportion of people in poverty in a middle-income neighborhood).

In order to release as much useful information as possible, statisticians must balance a number of factors. In particular, for Census 2000, the Bureau of the Census created weighting areas—geographic areas from which about two hundred or more long forms were completed—which are large enough to produce good quality estimates. If smaller weighting areas had been used, the confidence intervals around the estimates would have been significantly wider, rendering many estimates less useful due to their lower reliability.

The disadvantage of using weighting areas this large is that, for smaller geographic areas within them, the estimates of characteristics that are also reported on the short form will not match the counts reported in SF 1 or SF 2. Examples of these characteristics are the total number of people, the number of people reporting specific racial categories, and the number of housing units. The official values for items reported on the short form come from SF 1 and SF 2.

The differences between the long form estimates in SF 3 and values in SF 1 or SF 2 are particularly noticeable for the smallest places, tracts, and block groups. The long form estimates of total population and total housing units in SF 3 will, however, match the SF 1 and SF 2 counts for larger geographic areas such as counties and states, and will be essentially the same for medium and large cities.

This phenomenon also occurred for the 1990 Census, although in that case, the weighting areas included relatively small places. As a result, the long form estimates matched the short form counts for those places, but the confidence intervals around the estimates of characteristics collected only on the long form were often significantly wider (as a percentage of the estimate).

SF 1 gives exact numbers even for very small groups and areas; whereas, SF 3 gives estimates for small groups and areas such as tracts and small places that are less exact. The goal of SF 3 is to identify large differences among areas or large changes over time. Estimates for small areas and small population groups often do exhibit large changes from one census to the next, so having the capability to measure them is worthwhile.

August 2002

Summary File 3
Data Note 7

The following new section was added to Chapter 8, Accuracy of the Data.

CONSISTENCY WITH COMPLETE COUNTS

As described earlier, Census 2000 long form data were collected on a sample basis. Cities and incorporated places were used to determine sampling rates to support estimates for these areas. As a result, each city, incorporated place, school district, and county had addresses selected in the long form sample.

To produce estimates from the long form data, weighting was performed at the weighting area level. In forming weighting areas, trade-offs between reliability, consistency of the estimates, and complexity of the implementation were considered. The decision was made to form weighting areas consisting of small geographic areas with at least 400 sample persons (or about 200 or more completed long forms) that do not cross county boundaries. No other boundary constraints were imposed. Thus, total population estimates from the long form data will agree with census counts reported in SF 1 and SF 2 for the weighting area, county, and other higher geographic areas obtained by combining either weighting areas or counties. Differences between long form estimates of characteristics in the SF 3 and their corresponding values in the SF 1 or SF 2 are particularly noticeable for small places, tracts, and block groups. Examples of these characteristics are the total number of people, the number of people reporting specific racial categories, and the number of housing units. The official values for items reported on the short form come from SF 1 and SF 2.

Because the weighting areas were formed at a smaller geographic level, any differential nonresponse to long form questionnaires by demographic groups or geographical areas included in a weighting area may introduce differences in complete counts (SF 1 and SF 2) and the SF 3 total population estimates. Also, an insufficient number of sample cases in the weighting matrix cells could lead to differences in SF 1, SF 2, and SF 3 population totals. Thus, differences between the census and SF 3 counts are typical and expected.

In 1990, separate tabulations were not prepared for small areas below a certain size. In contrast, Census 2000 tabulations are being prepared for all areas to maximize data availability. This approach may lead to a greater number of anomalous results than what may have been observed with tabulations released from the 1990 census. A similar phenomenon occurred in the 1990 census when weighting areas respected city and place boundaries. Census counts differed from the long form data estimates in small places. As expected, these differences were sometimes large.

The SF 1 tables provide the official census count of the number of people in an area. The SF 3 tables provide estimates of the proportion of people with specific characteristics, such as occupation, disability, or educational attainment. The total number of people in the SF 3 table is provided for use as the denominator, or base, for these proportions. Estimates in the SF 3 tables give the best estimates of the proportion of people with a particular characteristic, but the census count is the official count of how many people are in the area.

The SF 1 gives exact numbers even for very small groups and areas; whereas, SF 3 gives estimates for small groups and areas, such as tracts and small places, that are less exact. The goal of SF 3 is to identify large differences among areas or large changes over time. Estimates for small areas and small population groups often exhibit large changes from one census to the next, so having the capability to measure them is worthwhile.

August 2002