

## Inconsistency Between Accuracy and Coverage Evaluation Revision II and Demographic Analysis Estimates for Children 0 to 9 Years of Age

Andrew Keller, U.S. Census Bureau, Washington, DC 20233

### ABSTRACT

The Census Bureau conducted the Accuracy and Coverage Evaluation Revision II (A.C.E. Revision II) with the goal of producing improved estimates of the net coverage of Census 2000. A.C.E. Revision II used dual system methodology to estimate the net coverage of Census 2000. Dual system estimates were created for population subgroups called post-strata. Post-strata groupings were based on race/Hispanic origin, tenure, relationship, household size, mail-back completion, area of residence, return rate, and age/sex. Population estimates were also created by Demographic Analysis, a separate and independent coverage evaluation program conducted at the Census Bureau. Demographic Analysis employed a macro-level approach for estimating undercount by comparing aggregate sets of administrative data while A.C.E. Revision II used a survey-based methodology.

The Census Bureau anticipated that A.C.E. Revision II estimates would be consistent with the Demographic Analysis estimates. However, A.C.E. Revision II estimated that Census 2000 had a small net overcount of children 0 to 9 years of age (although the estimate was not significantly different from zero) while Demographic Analysis estimated that Census 2000 had a net undercount of the same population<sup>1</sup>. Since the Demographic Analysis estimate for young children depended primarily on highly accurate recent birth registration data, the Demographic Analysis estimate is believed to be more accurate. This paper documents an analysis of the inconsistency between A.C.E. Revision II and Demographic Analysis estimates for children 0 to 9 years of age. In particular, this paper concentrates on the composition of the A.C.E. Revision II estimate. It offers an explanation of how the components of the A.C.E. Revision II estimate could have been modified to align A.C.E. Revision II

and Demographic Analysis estimates<sup>2</sup>.

### BACKGROUND

A.C.E. Revision II used two samples to evaluate coverage for Census 2000, the population sample (P sample) and the enumeration sample (E sample). The P sample assisted in measuring census omissions, persons that should have been enumerated in the census according to census residence rules but were not. The P sample consisted of people rostered from a sample of housing units (independent of the census) from a sample of census block clusters. It was populated based on the results from a person interview, independent from the census enumerations in the sample block clusters.

The E sample measured census erroneous enumerations, enumerations that should not have been included anywhere in the census or were included at the wrong location. The E sample consisted of people enumerated in the census from the same set of census block clusters selected for the P sample. E-sample enumerations who matched to P-sample people were counted as correct enumerations. Nonmatched E-sample people underwent a follow up interview to determine whether they were correctly enumerated.

A.C.E. Revision II divided the population into 7,456 crossed post-strata where smaller groupings were combined or collapsed to produce more stable estimates. A crossed post-stratum was a group of people sharing demographic and geographic characteristics that were assumed to have the same probabilities of inclusion in the census (U.S. Census Bureau 2004). A crossed post-stratum was composed of an E-sample post-stratum and P-sample post-stratum pair. Within a single crossed post-stratum, the dual system estimate (DSE) formula was defined as:

---

<sup>1</sup>For more information, see "Technical Assessment of A.C.E. Revision II," Decennial Statistical Studies Division (DSSD) A.C.E. Revision II Memorandum Series, Chapter PP-61. The corresponding web address is: <http://www.census.gov/dmd/www/pdf/ACETechAsses.pdf>

---

<sup>2</sup>This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

$$DSE_{ij} = census_{ij} * DDRATE_{ij} * \frac{CE_i / E_i}{M_j / P_j} \quad (1)$$

where:

*i* : the E-sample post stratum

*j* : the P-sample post stratum

*census<sub>ij</sub>* : the census count within a single crossed post-stratum (included "late additions" to the census, i.e., census records which were included that were too late for A.C.E. processing)

*DDRATE<sub>ij</sub>* : the ratio of data defined census records (excluding late additions) to all census records (including late additions) within a single crossed post-stratum<sup>3</sup>

*CE<sub>i</sub>* : weighted correct enumerations in E-sample post stratum *i*

*E<sub>i</sub>* : weighted E-sample enumerations in E-sample post stratum *i*

*M<sub>j</sub>* : weighted matches in P-sample post stratum *j*

*P<sub>j</sub>* : weighted P-sample records in P-sample post stratum *j*

Cumulative totals were compiled for multiple crossed post-strata for census counts and DSE population estimates:

$$census = \sum_i \sum_j census_{ij}$$

$$DSE_{PS} = \sum_i \sum_j DSE_{ij}$$

The subscript *PS* on the DSE term indicates that this DSE was calculated using post-stratification to generate estimates. Later in this paper, a different subscript on the DSE term is introduced for a DSE calculation which does not use post-stratification to form its estimate. Net coverage was defined for this dual system estimate and census count by:

$$coverage_{PS} = \frac{DSE_{PS} - census}{DSE_{PS}}$$

Positive values for net coverage imply that the census undercounted the population according to the dual system estimate. Conversely, negative values for net coverage imply that the census overcounted the population according to the dual system estimate.

For children 0 to 9 years of age, the dual system estimate of the population was 39,461,939. The census count of the population was 39,642,128. Hence, the net coverage using post-stratification was:

$$coverage_{PS} = \frac{DSE_{PS} - census}{DSE_{PS}}$$

$$coverage_{PS} = \frac{39,461,939 - 39,642,128}{39,461,939}$$

$$coverage_{PS} = -0.46\%$$

Similarly, the net coverage for the Demographic Analysis (DA) estimate was defined as:

$$coverage_{DA} = \frac{DA - census}{DA}$$

$$coverage_{DA} = \frac{40,684,311 - 39,642,128}{40,684,311}$$

$$coverage_{DA} = 2.56\%$$

This paper investigates why the A.C.E. Revision II and Demographic Analysis estimates for children 0 to 9 years of age were inconsistent. Consistency between A.C.E. Revision II and Demographic Analysis estimates was observed for all other age/sex groupings. To uncover why the dual system estimate indicated that the census overcounted the age 0-9 population, the components of the dual system estimate calculation were analyzed.

## COMPONENT EXPLANATION

From (1), four main components comprised every DSE for each crossed post-strata. First, a census total represented the final census count, including late additions to the census.

Second, a data defined rate, *DDRATE*, was defined for each crossed post-strata and contributed to the DSE. This data defined rate represented census records with sufficient information to be included as part of A.C.E. processing. The data defined rate was expressed as:

<sup>3</sup>In 2000, the census required two characteristics for a record to be data defined. Relationship, sex, race, Hispanic origin, and either age or year of birth counted towards the two necessary characteristics. A valid name also counted towards the minimum two characteristics. To be considered valid by the census, a name had to have at least three characters in the first and last name together. These data defined census records were eligible for A.C.E. processing.

$$DDRATE = \frac{\text{data defined census records}}{\text{total census records}}$$

The census records then selected to be part of the E sample were all data defined census records that received a classification of being a correct enumeration or an erroneous enumeration. Correct enumerations were eligible to have a matching P-sample person record.

Third, the correct enumeration (CE) rate quantified the ratio between total E-sample enumerations and a subset of correct E-sample enumerations. The CE rate was expressed as:

$$CERATE = \frac{CE}{E}$$

where:

*CE* : weighted correct E-sample enumerations

*E* : weighted total E-sample enumerations

Fourth, the match rate quantified the ratio between total P-sample people and a smaller subset of P-sample people who matched to a census enumeration. The match rate was expressed as:

$$MRATE = \frac{M}{P}$$

where:

*M* : weighted matching P-sample people

*P* : weighted total P-sample people

Each crossed post-stratum had a specific value for its census count, data defined rate, CE rate, and match rate. A data defined rate of 1.0 would signify that all census records were data defined within the crossed post-stratum. A CE rate of 1.0 would signify that all data defined census records within that E-sample post-stratum were correct enumerations. A match rate of 1.0 would signify that all P-sample records within that P-sample post-stratum were successfully matched to a census record.

The data defined rate, CE rate, and match rate were dependent on each other. This statement is best demonstrated through a simple example. Suppose that, for a given crossed post-stratum, a fixed number of data defined census records had been previously identified. Suppose, upon further examination, more census records were determined to be data defined. As a result, more census records were then eligible to become part of the E sample. Those new E-sample records could have been correct enumerations. As a result, the CE rate could have changed. Additionally, since matches could only occur between P-sample records and correct enumerations, the increased number

of correct enumerations would have given the possibility for an increase in the number of matches and the corresponding match rate. The following text proposes how these components comprising the coverage estimate for children 0 to 9 years of age may have been changed had the Census 2000 mail-return questionnaire<sup>4</sup> allowed for more demographic characteristics to be listed.

#### THE CENSUS 2000 MAIL-RETURN QUESTIONNAIRE

For Census 2000, respondents completing the census mail-return questionnaire were first asked to list the number of people living or staying at the housing unit on April 1, 2000. The mail-return questionnaire then had twelve roster spaces for listing people within the housing unit. To complete the first six roster spaces, the questionnaire asked for information regarding relationship to reference person, sex, age, date of birth, Hispanic origin, and race for those persons. The next six roster spaces on the questionnaire only requested name for those persons (persons seven through twelve). No information was requested concerning residents thirteen and beyond. Any housing unit with more than six persons was to be followed up by a census telephone-only operation to gather demographic information for persons seven and beyond.

For housing units with more than six residents, the lack of demographic information requested on the questionnaire for persons seven and beyond meant that these persons were non-data defined if more information was not obtained through followup. As a result, these non-data defined persons could not be part of A.C.E. processing. Consequently, it was hypothesized that children 0 to 9 years of age were more adversely affected by the length of the questionnaire than other age groups. If people on census forms were rostered oldest to youngest, this would be the case.

To study if the census questionnaire design had a detrimental effect on the coverage of children 0 to 9 years of age, it was first necessary to look at data defined rates of children 0 to 9 years of age compared to the other age groups:

---

<sup>4</sup>For the purposes of this study, the Census 2000 mail-return questionnaire refers to the questionnaire type which was either a) directly sent to the respondent through the mail or b) given to the respondent as part of the Update/Leave operation.

Table 1: Data Defined Rates Age 0-9 People Vs. People Age 10 or Older

Age	<i>DDRATE</i>
0-9	0.9568
10 or Older	0.9731

Census records which were not data defined could not be selected by the E sample. Nevertheless, non-data defined census records that could not be part of A.C.E. processing still influenced the computation of the DSE. So, how can it be determined if the lower data defined rates for children could in part be explained by the lack of roster space on the questionnaire? To do this, it was necessary to see if persons seven and beyond on census mail-return questionnaires were more likely to be children.

**METHODOLOGY**

For housing units with more than six people, persons seven and beyond who were non-data defined had demographic characteristics imputed. For Census 2000, 1,114,017 people had their characteristics imputed in housing units where six or more data defined people lived and at least one non-data defined person lived. It was thought that, for these people, characteristic imputations were necessary since six other people in the housing unit were data defined. Of these 1,114,017 people who had characteristic imputations, 452,751 (40.64%) were imputed as children 0 to 9 years of age. Comparatively, children 0 to 9 years of age represented only 14.49% of the total census count.

To see if persons seven and beyond on census mail-return questionnaires were more likely to be children, E-sample housing units with census forms indicating that they had more than six people were analyzed. Since the P-sample interview did not have a limitation in terms of rostering people, these E-sample housing units were compared to matching P-sample housing units. This comparison was done to get a snapshot of who actually resided at the housing unit and what the ages were for persons seven and beyond. Here is an example:

Table 2: Comparing Census and P-sample Responses in the Same Housing Unit

Number of census records at housing unit	Age (& Enumeration Status) of census records at housing unit		Number of P-sample people at housing unit	Age (& Match Status) on P-sample of P-sample people at housing unit	
	Age	Status		Age	Match Status
8 census records - 6 correct enumerations (CE) 2 imputations	41	CE	8 P-sample people - 6 matches (M) 2 non-matches (NM) with match code 'NC'	41	M
	36	CE		36	M
	10	CE		11	M
	9	CE		9	M
	6	CE		6	M
	4	CE		5	M
	11	Imputed		2	Non-matched NC
	6	Imputed		1	Non-matched NC

In this housing unit, the respondent listed eight people on the census mail-return questionnaire. However, only six census records were data defined. The last two

census records only had name information listed and could not have been data defined. As a result, the last two people were imputed into the census with ages of

eleven and six. Since the first six records matched it is reasonable to assume that demographic information for the two year-old and one year-old would have been listed on the census questionnaire if it would have been requested. Both the two year-old and one year-old received a match code of 'NC' for the matching process for A.C.E. Revision II. This code indicated that the P-sample person was not matched to the census person because only the name was collected by the census.

About 54% of the P-sample 'NC' cases in 2000 were children 0 to 9 years of age. However, only 40.64% of census records given characteristic imputations from housing units where six or more data defined people lived and at least one non-data defined person lived were imputed as children. Because its P-sample person interview did not limit the number of people collected from housing units, the A.C.E. Revision II results provide a more accurate picture of the makeup of these large housing units. As a result, the A.C.E. Revision II findings from the 'NC' cases demonstrate that more census imputations for children 0 to 9 years of age should have been made in housing units where six or more data defined people lived and at least one non-data defined person lived.

#### DSE CALCULATION - SINGLE CELL APPROACH

The next step is to study how the DSE would have been affected had the census form length not been an issue. Since demographic data was not collected for those 1,114,017 non-data defined census people who had their characteristics imputed in housing units where six or more data defined people lived and at least one non-data defined person lived, assume those people would have been data defined. Because A.C.E. Revision II results show that 54% of these imputations were children, that means  $1,114,017 * 54\% = 601,569$  children 0 to 9 years of age would have been data defined in 2000 had form length not been an issue.

From earlier, 452,751 children 0 to 9 years of age were imputed in housing units where six or more data defined people lived and at least one non-data defined person lived. As a result, 148,818 (601,569-452,751) more census records would be needed to make up for the discrepancy between 'NC' cases and census imputations. This new analysis also changes the *DDRATE* shown in Table 1.

$$DDRATE_{OLD} = \frac{\text{data defined census records (ddcr)}}{\text{census records (cr)}}$$

$$DDRATE_{OLD} = \frac{37,930,849}{39,642,128} = 0.9568$$

$$DDRATE_{NEW} = \frac{ddcr + \text{new ddcr}}{cr + \text{new cr}}$$

$$DDRATE_{NEW} = \frac{37,930,849 + 601,569}{39,642,128 + 148,818} = 0.9684$$

The newly modified census total (39,642,128 + 148,818 = 39,790,946) decreases the magnitude of the undercount for the Demographic Analysis estimate for children 0 to 9 years of age:

$$coverage_{DA} = \frac{DA - \text{census}}{DA}$$

$$coverage_{DA} = \frac{40,684,311 - (39,642,128 + 148,818)}{40,684,311}$$

$$coverage_{DA} = 2.20\% \text{ (vs. } 2.56\%)$$

With regard to A.C.E. Revision II estimates, adding the 148,818 new census records that were not previously census imputations presents a problem when re-analyzing the DSE. Since these 148,818 new census records did not have characteristics previously imputed, they have no demographic information. Since post-strata classification is based on demographic information, these new records cannot be assigned a post-strata. As a result, a DSE without regard to post-strata classification is used to reformulate the DSE. This manner of reformulating the DSE without using post-stratification is called the single-cell approach. This approach takes the total census count, data defined census count, correct enumeration rate, and match rate for all children 0 to 9 years of age and computes a new DSE estimate:

$$DSE_{single} = census_{0-9}$$

$$* DDRATE_{0-9}$$

$$* \frac{CE_{0-9}/E_{0-9}}$$

$$* \frac{M_{0-9}/P_{0-9}}$$

Here are the totals for all children 0 to 9 years of age used to compute the single-cell DSE estimate:

$$census_{0-9} = 39,642,128$$

$$DDRATE_{0-9} = \frac{ddcr}{tcr} = \frac{37,930,849}{39,642,128} = 0.9568$$

$$CE_{0-9} = 34,575,127$$

$$E_{0-9} = 36,389,165$$

$$M_{0-9} = 33,195,055$$

$$P_{0-9} = 36,506,208$$

As a result, the single-cell DSE estimate is:

$$\begin{aligned}
 DSE_{single} &= 39,642,128 \\
 * \frac{37,930,849}{39,642,128} \\
 * \frac{34,575,127}{33,195,055} &/ \frac{36,389,165}{36,506,208}
 \end{aligned}$$

$$DSE_{single} = 39,634,884$$

First, since the post-strata approach has been used to compare results between A.C.E. Revision II and Demographic Analysis, an adjustment must be made to account for the difference between the post-strata DSE estimate and the single-cell DSE estimate:

$$\begin{aligned}
 Adjustment_{PS} &= DSE_{PS} - DSE_{single} \\
 Adjustment_{PS} &= 39,461,939 - 39,634,884 = -172,945
 \end{aligned}$$

Comparing the two DSE estimates, 172,945 people are subtracted to adjust from the single-cell DSE estimate to the post-strata DSE estimate. As a result, when looking at revised DSE estimates under the single-cell approach, 172,945 people will be subtracted so a comparison can be made between revised A.C.E. Revision II results and Demographic Analysis results.

#### DSE COMPONENTS - REVISING CE AND MATCH RATES

As mentioned earlier, the data defined rate, CE rate, and match rate depend on each other. As a result, “creating” new data defined census records changes CE and match rates. Assumptions must be made as to how the new data defined people are classified with regard to correct enumerations and matches. The scenarios below propose different assumptions on how CE and match rates change because of the new data defined census records. Then, a new single-cell DSE, post-strata DSE, and net coverage estimate are determined based on those assumptions.

Scenario 1: Assume all new data defined census records are correct enumerations. Assume all P-sample person records coded as ‘NC’ non-matches are now matches.

From the analysis above, demographic information from 601,569 more children 0 to 9 years of age would have been collected had form length not been an issue. It is assumed that all of these children would have been correct enumerations. However, 601,569 is not added to the match total. In the P-sample, there were 256,211 weighted children with the match code of ‘NC’. As a

result, 256,211 is added to the match total. Because the ‘NC’ match code was thought to be under assigned since match codes could only be assigned at one stage of the matching process, adding 256,211 “new” matches to the original matching total provides a conservative lower limit for the number of matches to add when adjusting the original number of matches. The single-cell DSE estimate works out as follows:

$$\begin{aligned}
 DSE_{single} &= (39,642,128 + 148,818) \\
 * \frac{(37,930,849 + 601,569)}{(39,642,128 + 148,818)} \\
 * \frac{(34,575,127 + 601,569)}{(33,195,055 + 256,211)} &/ \frac{(36,389,165 + 601,569)}{36,506,208}
 \end{aligned}$$

$$DSE_{single} = 39,989,182$$

Subtracting 172,945 people to adjust from the single-cell DSE estimate to the post-strata DSE estimate:

$$\begin{aligned}
 DSE_{PS} &= DSE_{single} - 172,945 \\
 DSE_{PS} &= 39,989,182 - 172,945 = 39,816,237
 \end{aligned}$$

The net coverage estimate is then:

$$\begin{aligned}
 coverage_{PS} &= \frac{DSE_{PS} - census}{DSE_{PS}} \\
 coverage_{PS} &= \frac{39,816,237 - (39,642,128 + 148,818)}{39,816,237} \\
 coverage_{PS} &= 0.06\% \text{ (vs. } -0.46\%)
 \end{aligned}$$

With the assumption that all new data defined census records are correctly enumerated and that all P-sample person records coded as ‘NC’ non-matches are now matches, this revised DSE claims that the modified census total now slightly undercounts children 0 to 9 years of age. Although this new estimate is not significantly different from zero, it moves the original DSE estimate from a 0.46% overcount to a 0.06% undercount. This result also brings the dual system estimate for children 0 to 9 years of age closer to the Demographic Analysis estimate.

Scenario 2: Assume all new data defined census records are correctly enumerated matches.

Similar to Scenario 1, the 601,569 new data defined children 0 to 9 years of age are included as part of a revised DSE. However, for this scenario the assumption is that all these children would have matched to correct enumerations in the E sample. This higher assumption for the number of matches was made

because it was thought that the number of ‘NC’ cases was low since the match code could only be assigned at one stage of the matching process. Also, since weighting in the P sample is lower than weighting in the E sample, adding 601,569 “new matches” provides a generous upper limit to add when adjusting the original number of matches. The single-cell DSE estimate works out as follows:

$$\begin{aligned}
 DSE_{single} &= (39,642,128 + 148,818) \\
 &* \frac{(37,930,849 + 601,569)}{(39,642,128 + 148,818)} \\
 &* \frac{(34,575,127 + 601,569) / (36,389,165 + 601,569)}{(33,195,055 + 601,569) / 36,506,208}
 \end{aligned}$$

$$DSE_{single} = 39,580,544$$

Again, subtracting 172,945 people to adjust from the single-cell DSE estimate to the post-strata DSE estimate:

$$\begin{aligned}
 DSE_{PS} &= DSE_{single} - 172,945 \\
 DSE_{PS} &= 39,580,544 - 172,945 = 39,407,599
 \end{aligned}$$

The net coverage estimate is then:

$$\begin{aligned}
 coverage_{PS} &= \frac{DSE_{PS} - census}{DSE_{PS}} \\
 coverage_{PS} &= \frac{39,407,599 - (39,642,128 + 148,818)}{39,407,599} \\
 coverage_{PS} &= -0.97\% \text{ (vs. } -0.46\%)
 \end{aligned}$$

With the assumption that all new data defined census records are correctly enumerated matches, the revised DSE claims that modified census total provides a greater overcount than the original DSE post-strata estimate with the original census count. That is, the original DSE estimate is moved from a 0.46% overcount to a 0.97% overcount. This result also pulls the dual system estimate for children 0 to 9 years of age further away from the Demographic Analysis estimate.

The above scenarios represent upper and lower bounds to the DSE estimate respectively. Scenario 1 uses the actual P-sample weighted total for ‘NC’ non-matches that were missed because of form length and produces an upper bound to the DSE estimate. Scenario 2 assumes a larger number of ‘NC’ non-matches that were missed since the match code could only be assigned at one stage of the matching process. It produces a lower bound to the DSE estimate. In both scenarios however, the revised DSE does not make up enough ground to

assert that census form length was the sole cause for the discrepancy between A.C.E. Revision II and Demographic Analysis estimates for children 0 to 9 years of age.

### FUTURE WORK

Instead of reformulating the DSE using the single-cell approach, the 148,818 new census records that did not have characteristics previously imputed could be given demographic information using various assumptions. As a result, the new DSE would be calculated using post-stratification. The future work would focus on how to impute the demographic characteristics for these 148,818 children and how that would change the DSE calculations.

### CONCLUSION

This analysis explored whether the inconsistency between A.C.E. Revision II and Demographic Analysis estimates for children 0 to 9 years of age was because the Census 2000 mail-return questionnaire only allowed for demographic information from a maximum of six people. A questionnaire that would have allowed for more demographic information from more residents would have allowed for more census records and a higher data defined rate. However, the dearth of room on the questionnaire did not affect dual system estimates for children to the magnitude that A.C.E. Revision II and Demographic Analysis estimates for children 0 to 9 years of age would have been similar if form length had not been an issue.

### REFERENCES

Census 2000, “Accuracy and Coverage Evaluation of Census 2000: Design and Methodology,” U.S. Census Bureau, 2004.

Kostanich, D. (2003), “Technical Assessment of A.C.E. Revision II,” DSSD A.C.E. Revision II Memorandum Series, Chapter PP-61.