



UNITED STATES DEPARTMENT OF COMMERCE  
Economics and Statistics Administration  
U.S. Census Bureau  
Washington, DC 20233-0001


MASTER FILE

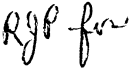
12/27/02

DSSD A.C.E. REVISION II MEMORANDUM SERIES # PP-10

PRED CENSUS AND SURVEY MEASUREMENT STAFF MEMORANDUM SERIES:  
CSM-A.C.E. REVISION II-04R

**MEMORANDUM FOR:** Donna L. Kostanich  
Chair, A.C.E. Revision II Planning and Management Group  
Assistant Division Chief, Sampling and Estimation  
Decennial Statistical Studies Division

**From:** Mary H. Mulry *signed 12/27/02*   
Chair, A.C.E. Revision II Quality Indicators Group  
Planning, Research and Evaluation Division

**Through:** David L. Hubble *signed 12/27/02*   
Assistant Division Chief, Evaluations  
Planning, Research and Evaluation Division

**Prepared By:** Anne T. Kearney  
Mathematical Statistician  
Planning, Research, and Evaluation Division

**Subject:** A.C.E. Revision II Missing Data Evaluation Study Plan

Attached is the A.C.E. Revision II missing data evaluation study plan. It is based on a report by Bruce Spencer, Report on Missing Data Evaluation, draft dated 10/22/02. Please direct any comments or questions to Anne Kearney, 301-457-4861.

cc: DSSD A.C.E. Revision II Memorandum Series Distribution List  
R. Killion  
D. Hubble

## **STUDY PLAN FOR MEASURING THE VARIANCE CONTRIBUTION FROM THE MISSING DATA PROCEDURES IN THE DUAL SYSTEM ESTIMATES FOR THE A.C.E. REVISION II**

### **I. BACKGROUND**

This project estimates the uncertainty in the A.C.E. Revision II dual system estimates (DSEs) due to choice of imputation model by drawing on the analysis of 128 reasonable alternatives to the imputation model conducted in 2001 (Keathley, et al., 2001; Kearney, et al., 2002; Keathley, et al., 2002). The ideal approach would be to repeat the very time-consuming analysis of reasonable alternatives for the A.C.E. Revision II estimator, but our limited resources do not permit it. Instead, we will develop an estimate of the additional variance due to the choice of imputation model by using the previous work for the original A.C.E (see Spencer, 2002a).

### **II. QUESTIONS TO BE ANSWERED**

How much variation is added to the DSEs as a result of the A.C.E. Revision II missing data methodology?

### **III. PROJECT DESCRIPTION AND METHODOLOGY**

#### **A. Objectives**

The purpose of this project is to estimate the uncertainty in the A.C.E. Revision II DSE as a result of the A.C.E. Revision II missing data methodology. Also, the bias adjusted vectors described in Section III.B, step 7 are input into the loss function for A.C.E. Revision II (see Spencer, 2002b).

#### **B. Statistical Methodologies**

In this section, we describe the creation of 128 vectors of coverage correction factors (CCFs) which serve as replicates. From these replicates, we estimate the variance from the missing data procedures. The methodology for the calculation of the CCFs adjusted for error due to missing data procedures are described in eight steps below (see Spencer, 2002a). We are able to accomplish this by using the 128 alternative combinations used in the evaluation of production missing data variance. From these replicates, we will estimate the variance from the missing data procedures. The variance estimation procedure is outlined in step 9 below. For a complete tasks list, see Section V. Additionally we describe comparisons we will do.

**Step 1.** Use the results for the previous 128 reasonable alternatives (see Keathley, et al., 2001), to the original missing-data methodology to calculate CCFs accounting for gross undercoverage (P-Sample match rate) and gross overcoverage (E-Sample correct enumeration rate) in each of

the new E-sample poststrata crossed by each of the new P-sample poststrata. (There are 7584 population groups when you cross the E-Sample poststrata with the P-Sample poststrata. Of the 7584 poststrata, 128 have zero entries.) For each alternative, construct a vector of CCFs. There will be 128 such vectors, one for each of the 128 reasonable alternatives, say  $\mathbf{y}_k$ ,  $1 \leq k \leq 128$  by the 7584 E-Sample crossed P-Sample poststrata.

**Step 2.** Compute the mean of the vectors, say  $\bar{\mathbf{y}}$ . Compute  $\mathbf{x}_k = \mathbf{y}_k - \bar{\mathbf{y}}$ .

**Step 3.** Multiply the deviations  $\mathbf{x}_k$  by a factor  $\phi$  to allow for the possibility that the original reasonable alternatives do not reflect sufficient variability. The default is to take  $\phi = 1$ . However, we think there is insufficient variability among the reasonable alternatives so we take  $\phi = 1.3$  (see Spencer, et al., (2002) for the rationale, and note that the empirical estimate of  $\phi$  is based on the 128 reasonable alternatives equally weighted, applied to the original poststratification.)

**Step 4.** Multiply the deviations  $\mathbf{x}_k$  by a factor  $\gamma$  to allow for the possibility that the imputation methods are improved relative to production. Note that  $\gamma$  only needs to reflect the ratio of variance of revised imputation methods to variance of production methods. If the Census Bureau had direct evidence, we could try to estimate  $\gamma$ , but since we do not have direct evidence we use  $\gamma = 1$ . This is a conservative estimate of  $\gamma$ .

**Step 5.** Pick a pair of alternative imputation treatments that will bracket the DSE, and refer to them as “high” and “low” alternatives. That is, we treat all unresolved matches as nonmatch (high) or match (low), etc. (Do not use alternative treatments for duplicates arising from the computer matching studies or for conflicting cases.) The alternative treatments are the same for original and A.C.E. Revision II DSE, except that the former requires adjustments to only the production level E- and P-sample files while the latter requires adjustments to both the production level files and the revision sample level files. To obtain high and low estimates, reset the following probabilities as detailed below.

	<u>High DSE</u>	<u>Low DSE</u>
Match Probability	0	1
Residence Probability	1	0
CE Probability	1	0

**Step 6.** Calculate the original DSE under the high and low treatments (using the original methodology and the original poststrata) and let the difference between the high DSE and low DSE be denoted by  $\delta$ . Similarly, calculate the A.C.E. Revision II DSE under the high and low treatments (under the A.C.E. Revision II methodology, applying the methods to the production and A.C.E. Revision II data files with the different E- and P-sample poststrata). Denote the difference between the high DSE and low DSE by  $\delta'$ . Let  $\eta = (\delta'/\delta)$ .

**Step 7.** Calculate the desired 128 replicates of CCFs as  $\mathbf{f}_{\text{impute}(k)} = \phi \times \gamma \times \eta \times \mathbf{x}_k + \bar{\mathbf{y}}$ ,  $1 \leq k \leq 128$ .

**Step 8.** Compute DSE estimates for each of the 128 vectors as  $\mathbf{f}_{\text{impute}}(\mathbf{k})^T * \mathbf{C} = \mathbf{D}_k$ ,  $1 \leq k \leq 128$  where C is the vector of Census counts.

**Step 9.** Calculate the variance in the DSEs as  $V(\mathbf{D}_k) = \Sigma(\mathbf{D}_k - \bar{\mathbf{D}})^2 / (k-1)$ .

### Comparisons

We will compare four different variances:

1. The standard deviation (531,751) based on the original missing data evaluation (See Keathley, et al., 2001).
2. The standard deviation of  $\mathbf{f}_{\text{impute}}(\mathbf{k}) * \mathbf{C}$ , i.e.,  $\sqrt{V(\mathbf{D}_k)}$ .
3. The standard deviation of the A.C.E. Revision II DSEs (from Doug Olson).
4. The standard deviation of the original production DSE.

Also we will compare the range of DSEs among the alternatives for the original missing data evaluation to the range among the A.C.E. Revision II missing data evaluation.

## **IV. DATA REQUIREMENTS**

### **A. Sources**

1. See Attachment 1 from Keathley (2002) for the list and locations of original 128 E-Sample and 128 P-Sample input files.
2. An estimation file for the production E-Sample and P-Sample that contain the A.C.E. Revision II poststrata codes.
3. The A.C.E. Revision II DSE estimation programs from Katie Bench.
4. The A.C.E. Revision II evaluation level missing data output files for P- and E-Sample person data, including A.C.E. Revision II poststrata for P- and E-Samples.
5. All A.C.E. Revision II poststrata level estimates need to calculate the DSEs.

### **B. Output**

1. See Attachment 2 for the output file names and layouts. These are alterations of the output from the production missing data evaluation as described in Keathley (2002). The DSE output files will be located on the UNIX at `/home/akearney/p1eval2k/reACE/`.

2. There are 7584 poststrata by 128 replicates (vectors) of CCFs. This matrix was transposed for input into the loss function (128 X 7584).

**V. DIVISION RESPONSIBILITIES**

After receiving the data files from DSSD, PRED took the steps outlined in Table 1.

**Table 1. TASK LIST FOR MISSING DATA EVALUATION FOR A.C.E. Revision II**

<b>Line</b>	<b>Task Name</b>
1	Draft Study Plan
2	Finalize Study Plan
3	Obtain person level file that contains poststrata for E- and P-Sample
4	Merge new poststrata onto 128 E- and 128 P-Sample files
5	Write DSE programs to get CCFs at new E and P poststrata for 128 combinations
6	Calculate prod. DSEs (CCFs) for 128 combinations with new poststrata
7	Write and run SAS programs to prepare input for “high” and “low” production DSEs
8	Calculate production DSEs for max and min with production DSE methodology
9	Obtain A.C.E. Revision II DSE programs
10	Write SAS programs to prepare input for “high” and “low” A.C.E. Revision II DSEs (including Revision Sample)
11	Run program from Step 10

Line	Task Name
12	Calculate A.C.E. Revision II DSEs for max and min
13	Decide which weighting scheme NOTE: we are using equal weights for the 128 alternatives
14	Which value do we use for $\phi$ ? Decided: $\phi = 1.3$
15	Which value do we use for $\gamma$ ? Decided: $\gamma = 1$ .
16	$\delta' = (\text{high} - \text{low})$ for reA.C.E. $\delta = (\text{high} - \text{low})$ for prod. A.C.E. $\eta = (\delta'/\delta)^2$
17	Calculate $f_{\text{impute}}(\mathbf{k})$ as $f_{\text{impute}(\mathbf{k})} = \phi \times \gamma \times \eta \times \mathbf{x}_k + \bar{\mathbf{y}}$ NOTE: we are not applying a factor based on duplication work.
18	QA Process 1. QA programs in steps 4,5,6,7, 8,10,17 2. Compare old and new national level DSEs for 128.
19	Processing Specification
20	Draft Report
21	Finalize Report

## VI. LIMITATIONS

As in the production evaluation of the missing data procedures, we are not including characteristic imputation alternatives in the A.C.E. Revision II estimate of variance due to missing data. Most of the variance due to characteristic imputation is accounted for in the estimation of sampling variance (see Kearney, 2002).

The variance calculated in this analysis is an approximation using alternatives from the evaluation of the production missing data system. Assumptions have been made in order to adjust the CCFs for the Revision Sample missing data procedures. For example, in step III.B.4, we are using  $G = 1$  which is conservative (may tend to over estimate the variance).

## VII. QUALITY ASSURANCE

The following tasks will have to be completed by PRED for the QA Process.

1. QA programs in tasks 4, 5, 6, 7, 8, 10, 12, 14, 17, 18 in Table 1 in Section V.
2. Compare old and new national level DSEs for 128.

## VIII. REFERENCES

Kearney, A.T. (2002), PLANNING, RESEARCH, AND EVALUATION DIVISION TXE/2010 MEMORANDUM SERIES: CM-MD-F-06 dated August 22, 2002, Reasons for not Considering Characteristic Imputation Alternatives in the Analysis of Missing Data Alternatives from Kearney for Documentation.

Kearney, A.T., Keathley, D.H., Belin, T.R., Petroni, R.J. (2002) "Alternatives of the A.C.E. Missing Data Evaluation," forthcoming *Proceedings of the 2002 Joint Meetings of the American Statistical Association, Survey Research Methods Section*.

Keathley, D. (2002) PLANNING, RESEARCH, AND EVALUATION DIVISION TXE/2010 MEMORANDUM SERIES: CM-MD-S-08 dated April 24, 2002, Program Documentation for the Calculation of Dual System Estimates and Covariance Matrices for the Missing Data Alternative Combinations in the Analysis of Missing Data for the Accuracy and Coverage Evaluation from Rita Petroni for Documentation.

Keathley, D., Belin, T., Bell, W., Kearney, A., Petroni, R. (2002) "Analysis of the Missing Data Alternatives for the 2000 A.C.E.," forthcoming *Proceedings of the 2002 Joint Meetings of the American Statistical Association, Survey Research Methods Section*.

Keathley, D., Kearney, A., Bell, W. (2001), paper for the Executive Steering Committee For A.C.E. Policy II, Report 12, ESCAP II: Analysis of Missing Data Alternatives for the Accuracy and Coverage Evaluation, dated October 11, 2001.

Mulry, M.H., A.C.E. Revision II Quality Indicators Chapter 7: Assessing the Estimates draft 8/21/02

Spencer, B.D. (2002a) Draft report, "Report on Missing Data Evaluation," October 22, 2002. Prepared by Abt Associates Inc. and Spencer Statistics, Inc. for the Bureau of the Census, Activity 20 - Deliverable 4, Task Number 46-YABC-7-00001, under contract no. 50-YABC-7-66020

Spencer, B.D. (2002b), Draft dated October 21, 2002, "Total Error Model and Loss Function Analysis for the A.C.E. Revision II Estimates of Population"

Spencer, B.D., Kearney, A.T., Keathley, D., Petroni, R., Belin, T., Mulry, M.H. (2002) Draft report dated August 1, 2002, Quantifying Bias from Missing Data Procedures in the 2000 A.C.E.

Table 1. Output Files from the 128 Missing Data Alternatives

Sample	Output Files <sup>1</sup>
P-Sample	alt1p1.dat - alt1p16.dat
	alt3p1.dat - alt3p16.dat
	alt4p#a.dat - alt4p#f.dat
	where # $\in$ { 1, 2, 4, 5, 7, 9, 12, 16}
	alt5p#a.dat - alt5p#f.dat
	where # $\in$ { 3, 6, 8, 10, 11, 13, 14, 15}
E-Sample	alt1e1.dat - alt1e16.dat
	alt3e1.dat - alt3e16.dat
	alt4e#a.dat - alt4e#f.dat
	where # $\in$ { 1, 2, 4, 5, 7, 9, 12, 16}
	alt5e#a.dat - alt5e#f.dat
	where # $\in$ { 3, 6, 8, 10, 11, 13, 14, 15}

<sup>1</sup> See Keathley (2002) for definitions of file names.

All of these files are on the UNIX at /home/keath001/p1eval2k/dse/apres24aout/.



**Table 2. Post-Stratum - Level DSE and CCF Output Files**

Row	Output Files
1	a101.dat - a116.dat
2	a301.dat - a316.dat
3	a4**1.dat - a4**6.dat where ** $\in$ {01, 02, 04, 05, 07, 09, 12, 16}
4	a5**1 - a5**6.dat where ** $\in$ {03, 06, 08, 10, 11, 13, 14, 15}

All of these files will be located on the UNIX at /home/akearney/p1eval2k/reACE/.

**Table 3. Layout for the Output Files in Table 2.**

Variable	Output From Row 1 and Row 2 (Table 2)		Output From Row 3 and Row 4 (Table 2)	
	Position	Format	Position	Format
Missing Data Alternative Combination Code	1-4	4.0 <sup>1</sup>	1-5	5.0 <sup>1</sup>
Post Stratum	6-9	4.0 <sup>1</sup>	7-10	4.0 <sup>1</sup>
DSE	11-23	13.2	12-24	13.2
CCF	25-34	10.8	26-35	10.8

<sup>1</sup> VPLX includes the decimal in the output for n.0 formatted variables