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

December 31, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES # PP-48

PRED CENSUS AND SURVEY MEASUREMENT STAFF MEMORANDUM SERIES: CSM-ACE-REVISION II-R2R

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 MHM
Chair, A.C.E. Revision II Quality Indicators Group Leader
Planning, Research and Evaluation Division

Through: David L. Hubble
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 Final Report

Attached is the A.C.E. Revision II missing data evaluation final report. It is based on a study plan, DSSD A.C.E. REVISION II MEMORANDUM SERIES # PP10, PRED CENSUS AND SURVEY MEASUREMENT STAFF MEMORANDUM SERIES: CSM-A.C.E. Revision II-R4R dated 12/27/02. Please direct any comments or questions to Anne Kearney, 301-763-6780.

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

A.C.E. Revision II

December 31, 2002

A.C.E. Revision II Missing Data Evaluation

Anne T. Kearney

Planning, Research, and Evaluation Division

USCENSUSBUREAU

Helping You Make Informed Decisions

Executive Summary

This project estimates the uncertainty in the Accuracy and Coverage Evaluation Revision II (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 for the original A.C.E.

The standard deviation of the 128 alternative DSEs of the A.C.E. Revision II missing data evaluation alternatives is 284,379, almost a 60 percent decrease in the standard deviation from the original A.C.E. evaluation (693,755). The original 128 alternatives were updated to make them comparable with the A.C.E. Revision II missing data evaluation. The A.C.E. Revision II missing data evaluation standard deviation is also lower than the standard errors from the original A.C.E. estimates (378,222) and the A.C.E. Revision II production estimates (541,631). This indicates that there is less non-sampling variability in the A.C.E. Revision II alternative DSEs compared to production A.C.E. Revision II.

1. 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 did not permit it. Instead, we developed 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, 2002).

2. METHODS

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 is described in eight steps below (see Spencer, 2002). We accomplish this by using the 128 alternative combinations used in the evaluation of production missing data variance. From these replicates, we estimate the variance from the missing data procedures. The variance estimation procedure is outlined in step 9 below. Additionally, we compare uncertainty statistics between the original missing data evaluation, the A.C.E. Revision II missing data evaluation, production A.C.E, and A.C.E. Revision II

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 $\mathbf{y}_k$, $1 \le k \le 128$ 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 ϕ 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 ϕ 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 γ to allow for the possibility that the imputation methods are improved relative to production. Note that γ only needs to reflect the ratio of variance of A.C.E. Revision II imputation methods to variance of production methods. If the

Census Bureau had direct evidence, we could try to estimate γ, but since we do not have direct evidence we use $\gamma = 1$. This is a conservative estimate of γ.

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, treat all unresolved matches as nonmatch (high) or match (low), etc. (We did 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 only to 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.

	High DSE	Low DSE
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 δ. 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 δ'. Let $\eta = (\delta'/\delta)$.

Step 7. Calculate the128 replicates of CCFs as $\mathbf{f}_{\mathbf{impute(k)}} = \phi\times\gamma\times\eta\times\mathbf{x}_{\mathbf{k}} + \bar{\mathbf{y}}$, $1 \le k \le 128$.

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

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

We use the variance calculated in step 9 to estimate the uncertainty in A.C.E. Revision II DSEs.

3. LIMITS

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 4 of Section 2, we are using G = 1 which is conservative (may tend to over estimate the variance).

4. RESULTS and CONCLUSIONS

4.1 The Uncertainty in the Alternative DSEs from the Missing Data Procedures in the A.C.E. Revision II

As outlined in Section 2, steps 1 through 9, we calculated the uncertainty due to the A.C.E. Revision II missing data procedures by making adjustments to the CCFs from the original 128 DSE alternatives. A factor in the adjustment was $\phi = 1.3$. Its purpose is to increase the variability among the original 128 alternatives. The value for ϕ was selected based on research by Spencer, et al. (2002). Another factor in the adjustment was $\eta = 0.410048$. Its purpose is to adjust for differences between the original and A.C.E. Revision II missing data methodology. The fact that the difference is smaller than one indicates a decrease in the variability of alternative DSEs for A.C.E. Revision II compared to the original production. The product of the two factors, ϕ x $\eta = 0.53$, is less than 1 resulting in an overall decrease in the variability of the 128 alternative estimates between A.C.E. Revision II and the original A.C.E. This outcome is reflected in Section 4.2 below.

4.2 Comparisons of Variation Due to Missing Data Procedures

In this section we compare the standard deviations calculated for the A.C.E. Revision II missing data evaluation to the standard deviation for the original missing data evaluation and to the original and A.C.E. Revision II production standard errors. We also compare the range of the DSEs from the A.C.E. Revision II missing data evaluation to the range from the original missing data evaluation.

Table 1 shows the estimates of error and the range of DSEs for production and for evaluations. The entry in the first row under the column heading “Production Std Error” is the sampling standard error from the production A.C.E. DSEs. The entry in the second row is the sampling standard error for the A.C.E. Revision II DSEs. The entry in the first row under the column heading “Evaluation Std Dev” is the standard deviation in the national level DSEs (summed across the original poststrata) for the 128 missing data alternatives from the production missing data evaluation adjusted to account for an underestimate in the spread of the DSEs (see Spencer, et al., 2002). The entry in the second row is the standard deviation in the national level DSEs (summed across the A.C.E. Revision II poststrata) for the 128 missing data alternatives from the production missing data evaluation adjusted for an underestimate of the spread in the DSEs and for the difference between the production and A.C.E. Revision II missing data methodology (Spencer, 2002).

Table 1. Original and A.C.E. Revision II Estimates of Standard Errors, Standard Deviations and Ranges of DSEs

	Production Std Error	Evaluation Std Dev	Evaluation Range of DSEs
Original A.C.E.	378,222	693,755[1]	3,417,035
A.C.E. Revision II	541,631	284,379	1,398,183

[1] The standard deviation in the original missing data evaluation (531,751) did not sufficiently reflect the variability in the DSEs from the choice of missing data methodology (Spencer, et al., 2002). We adjusted it so that the standard deviation from the A.C.E. Revision II missing data evaluation would be comparable.

The standard deviation of the original production missing data evaluation is based on the 128 reasonable alternatives, the original DSE estimation methodology, and the original 416 poststrata (adjusted due to a belief that there was not sufficient variability in the DSEs). We calculated the standard deviation to be 693,755. This is larger than the standard error of 378,222 in the production A.C.E. estimates indicating that non-sampling variability from the use of alternative missing data procedures was considerable at the national level.

The standard deviation of the 128 alternative DSEs of the A.C.E. Revision II missing data evaluation is 284,379, almost a 60 percent decrease in the standard deviation from the original production evaluation. It is also lower than the standard error from the original production estimates (378,222) and the A.C.E. Revision II production estimates (541,631). This indicates that there is less non-sampling variability in the A.C.E. Revision II alternative DSEs compared to production.

4.3 Comparison of the Average DSEs and the Range of DSEs

The A.C.E. Revision II evaluation average DSE is 276,998,074. The range of DSEs for the A.C.E. Revision II missing data evaluation is 1,398,183 with a maximum of 277,682,907 and a minimum of 276,284,723. This range of 1,398,183 is approximately 0.5 percent of the average DSE. The range of the national level DSEs for the original missing data evaluation (adjusted due to a belief that there was not sufficient variability in the DSEs) was 3,417,035. The decrease in the range between the original and A.C.E. Revision II evaluations further indicates that there is less non-sampling variability in the A.C.E. Revision II alternative DSEs compared to production.

5. 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., 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.

Spencer, B.D. (2002) 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., 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.