DIFFERENTIAL BIAS IMPACTS OF ALTERNATIVE CENSUS BUREAU HOT DECK PROCEDURES FOR IMPUTING MISSING CPS INCOME DATA

H. Lock Oh, Social Security Administration
Fritz Scheuren, Internal Revenue Service
Hal Nisselson, WESTAT

This paper compares 1973 Current Population Survey (CPS) money income data obtained from two alternative Census Bureau hot deck methodologies: the original procedure in use for the 1973 CPS sample and the revised (1976) procedure. Not only are these two imputation methods compared to each other but they are also compared to administrative data obtained in the Exact Match Study.

There are five sections to the paper. Some background on the Match Study is given in Section 1 because that Study forms the basis for our assessment of the original and revised procedures. These procedures themselves will not again be looked at in any detail since that has just been done in the companion paper (Oh and Scheuren, 1980a). Section 2 presents the basic comparisons upon which the evaluation results rest. Section 3 is a brief conclusion; Section 4 contains acknowledgements, footnotes and bibliographic references. Section 5 consists of the basic tables. (Section 5 is available upon request.)

1. BACKGROUND ON THE 1973 EXACT MATCH STUDY

The 1973 Exact Match Study was a joint undertaking of the Bureau of the Census and the Social Security Administration (SSA). Its starting point was the March 1973 Current Population Survey. A match was then made between the CPS sampled individuals and their social security benefit and earnings records. As part of this project, a limited set of tax items from 1972 Federal income tax returns were also furnished to the Bureau of the Census by the Internal Revenue Service (IRS) for matching to the CPS.

The 1973 effort represents a continuation by the Social Security Administration, the Bureau of the Census, and IRS for a long line of interagency data linkages for statistical purposes. For example, matching studies have been conducted to evaluate the last three decennial censuses and indeed are underway for 1980 as well (Bateman and Cowan, 1979). In surveys conducted by the Bureau of the Census for Social Security, interview schedules are combined routinely with administrative information on SSA earnings and benefits. Two-way matches involving IRS and SSA statistical samples have also been fairly common (Kilss and Scheuren, 1978).

Goals of 1973 Study and of Present Paper

The 1973 study was designed with a great number of specific goals in mind. High on the list of objectives was to evaluate and, potentially, to find ways of improving upon the procedures employed in carrying out the Current Population Survey. Specifically, we wished to evaluate:

1. procedures used to adjust for CPS coverage errors;

2. procedures used to adjust for CPS noninterview nonresponse;

3. CPS wage and property income reporting (by comparing it with the corresponding information provided to the IRS or to SSA);

4. CPS social security income reporting (by comparing it with SSA recorded amounts);

5. procedures used to impute for missing CPS income information.

The first two of these objectives, involving CPS weighting or estimation issues, have been addressed in a number of papers and reports, particularly, for example, in Scheuren, Oh, Yuskavage, and Vogel (1980). The middle two objectives, looking at response errors, have been extensively dealt with in a series of papers, most of which appear in Report No. 11 in the Social Security series, Studies from Interagency Data Linkages. The last objective, the primary focus here, has heretofore been addressed only in a preliminary way (Herriot and Spiers, 1975).

Our present evaluation of the CPS hot deck is similar to what was done by Herriot and Spiers. For example, we too rely heavily on administrative data. This evaluation differs, however, in that the intervening years have made it possible to improve the quality of the matching done in the Study. Also, because two sets of hot deck imputed values now exist for the March 1973 CPS, we can assess the relative importance of the sensitivity of the income distribution statistics to procedural changes made in arriving at the existing survey imputation system.

Comparing Administrative and Survey Sources

There are a number of conceptual differences which exist between survey and administrative reports of income. For example, Social Security benefit amounts are shown in SSA records on an accrual basis; the CPS concept is a cash one (Vaughan and Yuskavage, 1976). There are also important differences in the reporting of wage income. In Social Security records only covered wages are included and then only up to the taxable maximum ($9,000 for income year 1972).

Smaller differences exist between the CPS and IRS wage (and property income) concepts but they still are enough to raise interpretation issues when comparing administrative and survey incomes (Aziz, Kilss and Scheuren, 1978).

Because of conceptual issues, it is really not appropriate to consider CPS and administrative record differences as just measuring survey errors. It should also be noted that even if there were no conceptual differences, the administrative data is itself subject to error. For example, the tax return information matched as part of the study had not yet been subjected to audit. The Social Security information could change as well due to mispostings and late reporting. Nonetheless, it may be reasonable to contrast the relative differences between reported CPS incomes and selected administrative sources with similar differences between imputed CPS incomes and administrative sources. This is the approach taken here.

2. OVERALL SURVEY AND ADMINISTRATIVE INCOME DIFFERENCES

In this Section we describe the results obtained by making some limited comparisons between survey and administrative data. The basic approach we took went as follows:

1. Attention was confined solely to CPS families all of whose members had been properly matched to the Social Security and IRS records applicable for them. (Because the administrative data was not always on a person basis, i.e., tax units and claims often relate to two or more people, we were precluded from making comparisons for persons.)

2. The completely matched cases were reweighted so as to be representative of all U.S. families. This weighting adjustment was carried out in such a way that we "corrected" the survey not only for matching errors but also for undercoverage. 1/

3. CPS family survey income was defined in two ways: the income as originally reported or imputed and the revised income obtained by employing for the 1973 data new procedures in use since 1976.

4. CPS family administrative income was obtained by replacing survey wage, property and social security incomes with corresponding administrative source. Again, two versions of this concept were possible depending on whether the substitutions were made to the original or revised CPS. 2/

Notice that the CPS family income variables have several income types (public assistance, private pensions, etc.) that are unchanged by the administrative substitutions. This means, all other things equal, that we will understate the true differences which would exist if all the survey sources could have been replaced with administrative data. On the other hand, if we consider each of the substitutions separately, then there is some tendency, because of the nature of the approximations made, to overstate the real differences with the CPS. 3/ Any overstatement that might exist, however, is believed to be minor.

Overall Differences

Two overall summaries are provided which display some of the differences between the survey and administrative incomes of families in the March 1973 CPS. Table A contains mean family incomes for reported and imputed cases by source and type of income. Table B consists of the percentage distributions of total family income under the various alternative definitions.

Mean Family Income for Reported Cases.--For families all of whose income was reported, there are only small differences, due to editing changes between the original (1973) and revised (1976) processing systems. 4/ When the administrative amounts are substituted, mean incomes rise consistently, as might be expected, but only to a very modest degree -- except for property income where it appears there is a substantial degree of underreporting.

Figure 1.--Increase of Administrative Over Survey Income for Reported Cases

Income by Type	Percentage Increase	
	Original	Revised
Total.....	2.4	1.9
Wages.....	0.7	0.8
Social Security.....	3.6	4.6
Property.....	17.6	20.6
Other types.....	-	-

The contrast of the above with families where one or more of the income types was missing is quite marked.

Mean Family Income for Imputed Cases.--Families with imputed data differ in three major ways from those with all amounts reported. First, family incomes are considerably larger, by between 14.3% and 20.9% (depending on which measure is used). Second, the differences in the survey figures between the revised and original procedure are worth noting ($250 overall versus $77 for reporters). These were due mainly to the improved imputation methodology. 5/ The third difference is that when administrative data is substituted for survey amounts, the percentage increases tend to be a lot larger than for families who report all their income.

Figure 2.--Increase of Administrative Over Survey Income for Imputed Cases

Income by Type	Percentage Increase	
	Original	Revised
Total.....	8.3	5.8
Wages.....	4.2	2.8
Social Security.....	4.8	5.0
Property.....	81.4	68.1
Other types.....	-	-

The only exception to this is the case of social security benefits where the residual response and nonresponse biases, under the revised procedures, appear to be of about equal importance. 6/

Family Income Distribution Comparisons.--Table B presents family distribution data which is consistent with the kind of differences noted already in the means. As expected, the biggest discrepancies seem to lie in the upper tail of the percentage distributions, particularly the proportion of families with $50,000 or more. For example, when the administrative income concept is used, families with missing income data appear to be almost four times more likely to have incomes of $50,000 or more than do families reporting all income sources (i.e., 2.34% versus 0.65%). The original imputation procedure moreover only captures just over half of this difference (raising the percentage with incomes of $50,000 or more to 1.33%); with the revised imputation methodology, there is somewhat more improvement (to 1.57%) but unquestionably a substantial bias remains.

3. SOME CONCLUSIONS AND AREAS FOR FUTURE STUDY

This paper and the previous one at this session focus on components of the total mean square error of CPS income statistics: our first paper discusses the measurement of hot-deck variances. The present paper examines residual uncorrected biases that remain after imputation. From the results of both studies combined, it is possible for users of the CPS income data to determine the relative importance of bias and variance on the income distribution statistics for major subgroups of the population. In particular, some of the tables handed out at the Session should be of help in this regard.

What users of large data sets like the CPS need is a well-developed strategy for post-hoc survey adjustments (Scheuren, 1978). This paper, if it could be followed up by more work on nonresponse bias adjustments (as, for example, by Greenlees, Reece, and Zieschang, 1980) might form part of such an approach -- especially when combined with the techniques described below by Welniak and Coder (1980).

4. ACKNOWLEDGEMENTS, FOOTNOTES AND BIBLIOGRAPHIC REFERENCES

Acknowledgements

The computational work underlying this paper, and the previous one, was carried out by a number of staff members at the Bureau of the Census and the Social Security Administration (SSA). Chief among them were John Coder, Barry Fink and Emmett Spiers at Census; Faye Aziz, Henry Ezell and Linda Vogel at Social Security. Editorial assistance was provided by Beth Kilss and Wendy Alvey at the Internal Revenue Service (IRS). Typing help was given by Dawn Nester at IRS and Joan Reynolds at SSA.

Footnotes

1/ The Study weight used in these analyses was the "Combined Demographic and Administrative Weight Adjusted for Family Coverage," item 9.5 on the Supplementary Statistical Exact Match File. This supplementary file, available from Social Security on request contains all the variables employed in the current paper. Documentation for this file can be found in Report No. 10 of the Data Linkage Series. It should be noted that the weight being used does not adjust for the unknown number of illegal aliens eligible for interview in the March 1973 CPS (Lancaster and Scheuren, 1977). Weights which were calculated under alternative models of the number of illegal aliens are, however, included in the file and researchers may wish to use them as well.

2/ For wages, the process of replacing survey amounts with administrative ones went as follows: For married couples filing a joint return, IRS wages and salaries were substituted for the sum of the CPS wages of the couple. For nonmarried persons, married couples not filing jointly, or married persons not living with their spouses, matched IRS wages were substituted for the individual's CPS wages. In cases where no return was filed, the survey figure was retained unaltered. For social security benefits, the total amount of 1972 OASDI income from all claims of all beneficiaries in the CPS family was substituted for the corresponding survey amount. (Of the two ways of doing this shown on the Supplementary File, we chose to employ "Method I" in the present paper.) Property income was obtained by looking at each family member's administrative record to see if IRS dividends (after exclusion) were present; if they were, the amount shown was taken plus $200, or $100, depending on whether the return was joint, or nonjoint. The new dividend figure thus created was then added to any IRS interest present to create an IRS property amount. If CPS rental income was not indicated, then the IRS property amount was simply used to replace the CPS figure. If CPS rents were indicated, then the larger of the CPS or IRS property amount was to be taken. (This last step was employed because rental income or loss was not available on the IRS extract matched to the CPS.)

3/ There are conceptual problems raised in the handling of joint returns (when both spouses are not in the CPS family), social security benefits (when not all beneficiaries need be present), and property income (when there is a rental loss). For each of these cases, the approximation employed can lead to the administrative amount either including the income of a nonfamily member or of excluding losses that should be taken into account. We believe these problems to be minor because of the small frequencies with which such conceptual differences arise. Moreover, since we are using unaltered the CPS wage and property income of those who did not file tax returns the possibility of offsetting understatements must also be considered.

4/ See Consumer Income, P-60, No. 105,for a discussion of some of these changes and the impact they had in the 1975 CPS.

5/ For income year 1974, when the original and revised CPS processing procedures were first compared, the overall mean family income increase for reported and imputed cases combined was $209 or 1.4% [11]. (This figure, it might be added, is not strictly comparable to those in the present paper because for our work the family income amounts were truncated at $50,000 before the means were computed. Truncation was required because the public use version of the Exact Match data set was employed in the analyses.)

6/ These results for all CPS social security beneficiaries are roughly consistent with those obtained earlier in the paper by Herzog and Lancaster (1980). The higher mean benefits in Table A arise because all beneficiaries in the family are being considered, not just males 62 years or older.

References

[1] Aziz, F., Kilss, B. and Scheuren, F. (1978), 1973 Current Population Survey--Administrative Record Exact Match File Codebook, Part I--Code Counts and Item Definitions, in the Social Security series, Studies from Interagency Data Linkages, Report No. 9; Measuring the Impact on Family and Personal Income Statistics of Reporting Differences Between the Current Population Survey and Administrative Sources, in the Social Security series Studies from Interagency Data Linkages, Report No. 12.

[2] Bateman, D.V. and Cowan, C.D., "Plans for 1980 Census Coverage Evaluation," American Statistical Association Proceedings, Section on Survey Research Methods.

[3] Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1980), "Imputation of Missing Values When the Probability of Response Depends upon the Variable Being Imputed," submitted to the Journal of the American Statistical Association.

[4] Herriot, R. and Spiers, E. (1976), "Measuring the Impact on Income Statistics of Reporting Differences Between the Current Population Survey and Administrative Sources," American Statistical Association Proceedings, Social Statistics Section.

[5] Herzog, T.N. and Lancaster, C. (1980), "Multiple Imputation Modeling for Individual Social Security Benefit Amounts--Part I," American Statistical Association, Proceedings of the Section on Survey Research Methods.

[6] Herzog, T.N. (1980), "Multiple Imputation Modeling for Individual Social Security Benefit Amounts--Part II," American Statistical Association, Proceedings of the Section on Survey Research Methods.

[7] Kilss, B.K. and Scheuren, F.J. (1978), "The 1973 CPS-IRS-SSA Exact Match Study," Social Security Bulletin, Vol. 41, No. 10.

[8] Lancaster, D. and Scheuren F. (1977), "Counting the Uncountable Illegals: Some Initial Statistical Speculations Employing Capture-Recapture Techniques," American Statistical Association Proceedings, Social Statistics Section.

[9] Scheuren, F., Oh, H., Yuskavage, R., and Vogel, L. (1980), Methods of Estimation for the 1973 Exact Match Study, Report No. 10 in the Social Security series, Studies from Interagency Data Linkages.

[10] Scheuren, F.J. (1978), "Discussion of the Paper Entitled 'Income Data Collection and Processing for the March Income Supplement to the Current Population Survey," by J. Coder," Proceedings of the Data Processing Workshop: Survey of Income and Program Participation, U.S. Department of Health, Education, and Welfare.

[11] U.S. Bureau of the Census, Consumer Income, Current Population Report Series P-60, Nos. 33, 68, 75, 103, 105. See also the related reports in the Bureau's Technical Paper series (Nos. 8, 17 and 35).

[12] Vaughan, D. and Yuskavage, R. (1976), "Investigating Discrepancies Between Social Security Administration and Current Population Survey Benefit Data for 1972," American Statistical Association Proceedings, Social Statistics Section.

TABLE A.--MEAN FAMILY INCOME IN 1972: REPORTED AND IMPUTED CASES BY INCOME TYPE, SOURCE OF DATA

(IN DOLLARS)

INCOME TYPE	REPORTED CASES				IMPUTED CASES			
	SURVEY INCOME		ADMIN. INCOME		SURVEY INCOME		ADMIN. INCOME	
	ORIGINAL	REVISED	ORIGINAL	REVISED	ORIGINAL	REVISED	ORIGINAL	REVISED
TOTAL FAMILY INCOME	12,259	12,336	12,558	12,566	14,009	14,259	15,178	15,082
WAGE AND SALARY	11,488	11,460	11,571	11,555	12,399	12,443	12,923	12,795
SELF-EMPLOYMENT INCOME (NONFARM)	7,176	6,600	7,176	6,600	10,953	9,462	10,953	9,462
SELF-EMPLOYMENT INCOME (FARM)	4,290	4,021	4,290	4,021	4,785	4,251	4,785	4,251
PROPERTY INCOME	1,035	1,057	1,217	1,275	1,212	1,425	2,313	2,396
SOCIAL SECURITY/RAILROAD RETIREMENT	2,353	2,328	2,437	2,435	2,418	2,431	2,533	2,553
PUBLIC ASSISTANCE	1,717	1,672	1,717	1,672	1,395	1,347	1,395	1,347
OTHER GOVERNMENT TRANSFER	1,719	1,735	1,719	1,735	1,658	1,890	1,658	1,890
OTHER INCOME	1,916	1,911	1,916	1,911	2,016	2,335	2,016	2,335

Note: Average farm and nonfarm self-employment income decreases from the original to the revised procedure. This is primarily because the number of families with such incomes increased faster than the income amounts received. Amounts were all truncated at $50,000 before the means were computed. This has the effect, as Table B indicates of understating somewhat the differences brought about by the introduction of the administrative income data.

TABLE B.--FAMILY INCOME SIZE PERCENTAGE DISTRIBUTIONS IN 1972: REPORTED AND IMPUTED CASES BY SOURCE OF DATA

INCOME SIZE CLASSES (IN DOLLARS)	REPORTED CASES				IMPUTED CASES			
	SURVEY INCOME		ADMIN. INCOME		SURVEY INCOME		ADMIN. INCOME	
	ORIGINAL	REVISED	ORIGINAL	REVISED	ORIGINAL	REVISED	ORIGINAL	REVISED
TOTAL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
UNDER 5,000	17.09	17.28	16.92	17.00	14.31	13.95	13.12	14.40
5,000 TO 6,999	10.35	10.40	9.98	10.09	9.31	9.05	8.51	8.60
7,000 TO 9,999	16.90	16.90	16.22	16.34	15.53	14.69	13.70	13.85
10,000 TO 12,999	17.14	17.46	16.39	16.82	16.14	14.76	14.58	14.99
13,000 TO 14,999	8.93	8.99	9.77	9.81	8.35	8.90	9.06	7.39
15,000 TO 17,999	10.76	10.69	10.79	10.77	11.11	11.24	11.17	11.49
18,000 TO 19,999	4.96	4.99	5.16	5.14	5.13	5.66	5.80	5.77
20,000 TO 22,499	4.60	4.47	4.53	4.37	5.29	5.52	5.45	5.63
22,500 TO 24,999	2.75	2.65	2.89	2.82	2.87	3.95	3.84	3.68
25,000 TO 29,999	3.24	3.14	3.61	3.41	4.68	4.46	5.84	5.16
30,000 TO 49,999	2.66	2.47	3.01	2.77	5.96	6.24	6.57	6.70
50,000 OR MORE	.63	.55	.74	.65	1.33	1.57	2.37	2.34