ESTIMATING THE VARIANCE IMPACT OF MISSING CPS INCOME DATA

H. Lock Oh, Social Security Administration
Frederick J. Scheuren, Internal Revenue Service

This paper describes an experimental study designed to estimate the imputation variance arising from the use of the Census Bureau's existing hot deck procedures in the March Income Supplement to the Current Population Survey (CPS). We have employed the multiple imputation methodology of Rubin (1978) to estimate this variance. Our focus is on the impact on statistical inference of the present CPS variance estimation procedures which ignore the item nonresponse problem and, hence, lead to confidence intervals which are too short. (Scheuren, 1975)

There are seven main sections to the paper. In section 1 we set the stage by providing a brief historical background of changes in CPS hot deck methodology since 1962, when the hot deck was first introduced into the income supplement processing. Some theoretical considerations in the estimation of the hot deck variance are taken up in section 2. Section 3 describes the results of the multiple imputation application we attempted with the March 1978 CPS. Areas for further study are suggested in section 4 which forms a bridge to our companion paper (Oh and Scheuren, 1980b) on the impact of the CPS hot deck on reducing income nonresponse biases. The remaining three sections contain appendix material: acknowledgements, and selected hot deck references (section 5), some details on the derivation of the multiple imputation variance formulas used (section 6), and the basic tables (section 7). (Sections 6 and 7 are available upon request.)

1. HISTORICAL BACKGROUND ON THE CPS HOT DECK

Hot decks may be divided in general into two types: static and dynamic. The traditional hot deck methodology, as in the CPS, is based on a static system of categorical cross-classifications which are chosen in advance of the data collection. It must be possible to array both respondents and nonrespondents on all the dimensions of the "hot deck matrix" as it is sometimes called. The newer "hot decks" create the matrix dynamically. That is, for a given nonrespondent, a cell is defined of respondents similar in some specified way, i.e., "nearby" respondents in terms of a distance measure that incorporates both categorical and continuous variables (Colledge et al, 1978; Sande, 1979a). The matching (or cell defining stage) in such a system takes full advantage of the random access computer technology which is now widespread. A great deal of theoretical and empirical work appears to be going on with dynamic hot decks (e.g. Ernst, 1980). While not the topic for today, it does seem that the next major advance in the CPS will be in this area.

Our historical description of the CPS income hot deck will be divided into four parts: First there will be a simple illustration of a static hot deck, the remainder of the presentation will examine in turn each of the major advances made in the CPS since 1962.

Static (CPS-like) Hot Deck

In order to make clear some of the main issues in static hot decks figure 1 is provided below.

Figure 1.--Income Hot Deck by Age and Sex

	Sex	
	Male	Female
Under 65 years of age	$18,764 $9,832 A 1	None C 1
65 years of age or older	$3,623 B 1 2	$2,468 D None

In each of the four cells, labelled A to D, the observations have been divided into two parts: in the upper left half of each cell income amounts are shown for respondents; in the lower right half there is an indicator of how many nonrespondents there were in the particular age-sex subgroup defined by the hot deck "matrix". Perhaps by way of introducing some of the continuing issues in hot deck technology each of the cells warrants discussion separately:

Cell A.--In this cell there is more than one respondent and hence the issue of how to assign respondents to nonrespondents arises. In many of the older hot decks the sequence in which the computer file (or card deck) was processed determined the assignments (Bailar and Bailar, 1978). More control over this "randomization" step has been considered desirable for a long time (Naus, 1975; Scheuren, 1975) and a systematic sample is now taken in the CPS income supplement processing (Coder, 1978). Other sampling schemes (Scheuren, 1980) have also been proposed .

Cell B.--For this cell there are more nonrespondents than respondents. We thus are brought face-to-face with one of the ways that the hot deck can increase the variance of our estimators. (The "randomization" step mentioned above is another.) Controlling the number of times a respondent is used as a "donor" remains a serious concern and may best be resolvable in the context of dynamic hot decks where "penalties" can be assigned after a respondent is used, reducing the chance of its reuse (Ford, 1980).

Cell C.--In this cell there are no respondents to assign income to the nonrespondent. Since this is a common occurrence for static hot decks employing many matching variables, typically special procedures are set up to take care of such cases. (In the latest CPS approach, for example, there are twenty-four stages that nonrespondents can "fall through" until they hit a cell with at least one respondent.)

Cell D.--This cell does not have any nonrespondents but does have respondents. Now, by the very nature of the static hot deck technique, there is no way this cell can contribute to the imputation of missing income data. Unlike a regression or log-linear modelling approach to imputation (Herzog and Lancaster, 1980) the hot deck does no smoothing of the respondent information before assigning the missing data. This is a distinct disadvantage and there have been proposals to modify the hot deck so it incorporates some smoothing (Scheuren, 1975-76; Schieber, 1978).

Some of the points not brought out by figure 1 but which we also will be concerned with are the fact that not only is the hot deck a variance increasing procedure but also, usually, the imputed data is treated as if it were reported, leading to variance estimates which can badly understate the actual variance. In general there is considerable evidence which suggests that residual nonresponse biases remain a serious problem in the CPS despite many improvements over the years (Oh and Scheuren, 1980b). Finally, the hot deck does not always adequately preserve relationships between the imputed and reported information (Welniak and Coder, 1980).

Early CPS Hot Deck Methods, 1962-1965

The Census Bureau has collected annual income information as part of the CPS since 1947. (See, for example, Technical Paper No. 35.) Perhaps because of the modest number of income questions employed, item nonresponse seems to have been only a minor problem in the early years of the CPS (Ono and Miller, 1969). Initially, percentage distributions by income level in published Bureau reports were based solely on those cases which reported completely. (The assumption implicit in this method was that persons who do not provide income information have the same income distributions as those who do.)

Because of a concern about the bias that the income nonresponse could introduce, a comparison was made, using income data for 1958, of the income distributions obtained before and after the imputation of income to nonrespondents on the basis of their known demographic and economic characteristics. This comparison indicated that the procedure for making individual assignments of income to nonrespondents resulted in slightly higher estimates of the proportion of families and individuals in the upper income classes than were obtained from the distributions based solely on those reporting on income. (Consumer Income, Series P-60, No. 33.)

The experimental work with the 1958 CPS lead, in the March 1962 survey, to the publication by the Census Bureau of income statistics where nonrespondents on income were imputed the reported income of a person with similar demographic and economic characteristics. A hot deck approach was used in which respondents and nonrespondents were classified by age, sex, family status, color, weeks worked and major occupation group. The income amount assigned to a nonrespondent was that observed for another person with the same characteristics selected systematically in the order in which the records were processed.

With minor variations the imputation procedure employed for March 1962 was used until March 1966. Comparison with administrative data from the 1963 Link Study (Scheuren, Oh and Alvey, 1980) suggests that important refinements were needed in these initial attempts. Nonetheless, relative to the alternative of not imputing at all, the evidence also indicates that it was better to impute for the missing data than otherwise, provided one is interested solely (as was true for the Consumer Income reports published at that time) in income size distribution questions. This is to be contrasted with "classificatory" questions, such as the number of persons in poverty, which appear (e.g., Consumer Income, P-60, No. 68, or Scheuren, 1970) to have been very sensitive to the imputation process.

Improved Hot Deck Methods, 1966-1975

Starting with the March 1966 CPS the Bureau made a series of important changes which, by March 1968, substantially improved the initial method of imputation. For example, with the 1962 procedure, if an interviewed person did not answer one or more of the income items, all of them were imputed. Beginning with the March 1966 survey, however, in the event a respondent did not answer one or more of the income questions, the missing income data for this person were imputed for only those income items which were not answered.

The main feature of the improved procedure, distinguishing it from earlier approaches, was a more refined method for imputing missing income data which expands the use of information already known about that person. Among the major improvements made affecting the income data were the following: (1) An expanded set of social and economic characteristics within which the imputations are made (in particular more detail by age, race, occupation, weeks worked, sex and type of family membership); and (2) the elimination of inconsistent reporting which resulted in having workers with no earnings and earners with no weeks worked. The improved imputation procedures also assigned missing earnings entries first and then utilized the earnings in the assignment of other income. For more details see Consumer Income, P-60, No. 75 or Spiers and Knott (1969).

Current Hot Deck Methods, 1976 to Date

The last major revision in the CPS imputation of income was designed for use in the March 1976 survey (Coder, 1978). There were two main objectives:

1. In an attempt to further reduce the nonresponse bias the demographic and economic characteristics used in the hot deck were greatly enlarged.
2. The second objective was to maintain--wherever possible--observed relationships for respondents among income, work experience, and longest job variables in imputing missing information for nonrespondents.

To reduce the nonresponse bias the following major changes were instituted in the demographic and economic characteristics used to define nonrespondents as "similar" (Consumer Income, P-60, No. 103):

1. Missing Earnings--Earnings are now imputed using a greater number of demographic and economic characteristics than previously. These added characteristics include educational attainment, labor force status of spouse, marital status, number of children, region, and type of residence (e.g., inside metropolitan areas of 1,000,000 population or more, farm residence, etc.). Characteristics used in the previous imputation system but expanded to provide more detail in the new system include age, family relationship, occupation of longest job, class of worker of longest job, weeks worked, full-time/part-time work status, and race-ethnic origin (Spanish origin included).
2. Missing Other Income--Expanded detail and numbers of characteristics were also used for the imputation of the "unearned" sources of income such as Social Security benefits, public assistance, unemployment compensation, dividends, etc. Characteristics added for use in the new imputation system for "unearned" income include years of school completed, weeks worked, reason for not working (nonworkers), marital status, number of children, total family earnings, region, type of residence, and, when available, the reported recipiency pattern for the "unearned" sources of income. Characteristics used in the previous imputation system but expanded to more detail in the new one include age, family relationship, amount of earnings, and race-ethnic origin (Spanish origin included).

To achieve the second objective (the preservation of relationships among income, work experience, and longest job categories), the imputation system was designed to impute all missing information to a nonrespondent from the same "similar" respondent. The addition of current labor force information, and for married persons--current labor force status of spouse, also helps to preserve relationships between "current" labor force status and job and "last year's" work experience, longest job, and earnings for individuals within the family. The previous procedure imputed missing work experience, longest job, earnings, and unearned income in separate stages. Although the revised procedure attempts to impute earnings, work experience and longest job in a single stage, a second stage is still required to impute the "unearned" sources. Several revisions were made to the imputation procedure to help preserve the distribution of "unearned" income sources for individuals. First, the reported recipiency pattern for "unearned" income is always used--when available--to impute missing dollar amounts. Second, husbands and wives are imputed missing "unearned" income information as a unit in order to prevent inconsistencies between the amounts and sources imputed to the couple.

2. DERIVATION OF SIMPLIFIED VARIANCE ESTIMATORS FOR THE CPS HOT DECK

In this section we will briefly describe some of the considerations that went into our experiments in estimating the imputation variance arising from the use of the revised CPS hot deck. The presentation is divided into two parts. First we set forth the basic "post-stratification" theory appropriate to a static hot deck. The multiple imputation estimation of some of these variance components is then described. (We leave to the end of this paper a few comments on the practical implementation issues in routinely estimating the CPS imputation variance.)

Basic "Post-Stratification" Theory

Static hot decks, like those employed in the CPS, bear a close relationship to weighting class (or post-stratification) adjustments in sample surveys (Oh and Scheuren, 1980c). In order to flesh out this general observation we will make the following simplifying assumptions:

1. All the imputations can be done in one stage.
2. We have a simple random sample.
3. There are no nonsampling errors other than nonresponse errors.
4. The imputation procedure is unconditionally unbiased.
5. The probability of a nonresponse is independent for each unit in the same hot deck cell.

Anyway, suppose we have a hot deck with h = 1, ..., H cells where--

$\underline{m}=(m_1,\dots,m_h,\dots,m_H)$ is the number of respondents in each cell ($m_h > 0$ for all h)

$\underline{n}=(n_1,\dots,n_h,\dots,n_H)$ is the total sample in each cell and $\sum n_h = n$

$N=(N_1,\dots,N_h,\dots,N_H)$ is the total (unknown) population in each cell and $\sum N_h = N$.

Further let $\{Y_{hi}\}$ be an income amount to be observed for the i^{th} individual, i = 1, ..., N_h, in the h^{th} cell where $\bar{Y}_h$ and $\bar{Y}$ are the cell and overall population means respectively. Finally, we will denote by $\{V_h\}$ the within cell variances given by

$$(N_h-1)V_h = \sum_{i=1}^{N_h} (Y_{hi}-\bar{Y}_h)^2 .$$

Now it can be shown that if the imputation is carried out independently within each cell the mean square error on the estimated total $\tilde{Y}_I$, conditional on $\underline{n}$ and $\underline{m}$, can be written as the sum of four components:

1. The unconditional bias (squared) arising from the fact that the respondents and nonrespondents differ within the hot deck cells in the distribution of the $\{Y_{hi}\}$;
2. The conditional bias (squared) arising from the fact that our estimate of each subpopulation or cell size $\{n_h \ N/n\}$ will differ in any particular sample from N_h even though they are equal on the average;
3. The conditional variance of a simple ratio adjustment for the nonresponse within each cell obtained by weighting the respondents by the quantity n_h/m_h; and
4. The variance component due to the hot deck procedure of choosing a (single) respondent's data for each nonrespondent. (This is the only variance component that can be reduced by multiple imputations.)

In this presentation we will only consider the last three components of the mean square error--reserving for our companion paper the measurement of the unconditional bias of the CPS hot deck. For the imputed items $\{Y_{hi}\}$, if there is no unconditional bias, then we can write the conditional MSE$(\tilde{Y}_I)$ as MSE$(\tilde{Y}_I|\underline{n},\underline{m})$ =

$$\left\{\text{Bias}(\tilde{Y}_I|\underline{n},\underline{m})\right\}^2 + \text{Var}(\tilde{Y}_I|\underline{n},\underline{m}) \text{ where}$$

$$\text{Bias } (\tilde{Y}_I|\underline{n},\underline{m}) = \sum_{h=1}^{H} (\bar{Y}_h-\bar{Y}) \left(N_h-n_{hn}\frac{N}{}\right)$$

and

$$E\left\{\text{Bias } (\tilde{Y}_I|\underline{n},\underline{m})\right\} = 0 \text{ .}$$

The conditional variance $\text{Var}(\tilde{Y}_I \mid \underline{n},\underline{m})$ can be divided into two components: the variance of the weighting class estimator, $\tilde{Y}_R = \frac{N}{n} \sum_{h=1}^{H} n_h \ \bar{y}_h$ that is

$$\text{Var}(\tilde{Y}_R|\underline{n},\underline{m}) = \sum_{h=1}^{H}\left(\frac{Nn_h}{n}\right)^2 \left(1- \frac{m_h}{N_h}\right)\frac{V_h}{m_h}$$

where the $\left\{\bar{y}_h\right\}$ are the sample means of the respondents, plus an additional component whose magnitude depends on the particular way in which the hot deck assignments are made.

Hansen, Hurwitz and Madow (1953) show that if the number of nonrespondents is less than the number of respondents, and a simple random subsample is drawn without replacement, then the relative increase in the variance over a simple ratio adjustment will be at most 12.5% within any one cell provided $(N_h-n_h)/N_h \doteq 1$. Further results in the literature are given for sequential selection under serial correlation models (Bailar and Bailar, 1978), integerization methods (Oh and Scheuren, 1980c), and several other techniques (Ernst, 1980).

The present CPS hot deck procedure uses a systematic sampling technique at each stage taking the donor respondents without replacement to the extent possible. A random permutation is formed of the respondents who are then systematically assigned to the nonrespondents in the same cell. Should there be more nonrespondents than respondents, that is should $n_h-m_h > m_h$, then the assignments are made by repeatedly cycling completely through the m_h respondents until all the imputations have been made. For instance, suppose $n_h-m_h = km_h + m_h^*$ where k is an integer and $m_h^* < m_h$, then we would choose $n_h-m_h^*$ of the respondents exactly k times and the remaining m_h^* respondents (k+1) times. Under these circumstances the relative increase in the variance of the CPS hot deck procedure over the variance of a simple inflation estimator would, <u>for a particular cell h</u>, be approximately

$$\frac{\text{Var}(\tilde{Y}_I)-\text{Var}(\tilde{Y}_R)}{\text{Var}(\tilde{Y}_R)} \doteq \frac{x-x^2}{(k+x)^2}$$

where $x_h=m_h^*/m_h$, N_h is assumed large relative to m_h, and we ignore the complex multi-stage cluster design of the CPS (treating it as if it were a simple random sample). It might be worth noting that the relative variance increase is a maximum at k = 1, where it can be as large as 12.5% (as mentioned above); the relative variance increase declines quickly however: for k = 2 it can be no more than 4.2%, for k = 3, 2.1% and so on. (Oh and Scheuren, 1980c)

<u>Multiple Imputation Estimation Approach</u>

One way to try to estimate the components of the MSE $(\tilde{Y}_I)$ is to employ the multiple imputation approach advocated by Rubin (1978, 1980). In the context of the CPS hot deck this would mean that, conditional on the cross-classification scheme, an independent selection of donor-respondents would have to take place at each stage in the assignment process. For cost reasons we had to confine ourselves just to two independent imputations. There were also a number of other factors which make the estimates only rough indications of the impact of the imputations on the CPS income estimates:

1. A sizeable fraction of the CPS nonrespondents were imputed in cells where there was only one respondent (roughly one-third of the imputations were of this type); hence, in a multiple imputation context we would get the same respondent in both cases (leading to an underestimate of the variance if uncorrected).
2. Because the imputations were carried out independently and with replacement there were also a number of cases where the first and second imputations were the same even in cells with more than one respondent. This too had to be taken account of in our variance calculations.

3. The present hot deck procedure does not provide the end user with a way of determining which hot deck cell was employed in the imputations; hence, we were unable to explicitly condition on $\underline{n}$ and $\underline{m}$ because the actual $\left\{n_h\right\}$ and $\left\{m_h\right\}$ are unknown to the analyst. We have however implicitly conditioned on these quantities with the result that at best we are able to estimate just two of the three variance components. (The squared conditional bias term is not being estimated.)

In the basic results described in the next section for the March 1977 CPS, attention was confined to overall income distribution proportions $\left\{\tilde{\tilde{P}}_j\right\}$ in each of j=1,...,J standard income size classes. For these statistics the variance was estimated by the expression

$$\widetilde{\dot{V}}ar(\tilde{\tilde{P}}_j) = \frac{\tilde{P}^R_j(1-\tilde{P}^R_j)}{m} + \frac{3}{2}\left(\frac{n-m}{n}\right)^2 (\tilde{P}^N_{j1}-\tilde{P}^N_{j2})^2$$

where $\tilde{P}^R_j$ is the proportion of respondents in the jth income class, $\tilde{P}^N_{j1}$ is the proportion of nonrespondents in the j^{th} class in the first imputation, $\tilde{P}^N_{j2}$ is the proportion of nonrespondents in the j^{th} class in the second imputation. As Section 7 makes clear, $Var(\tilde{\tilde{P}}_j)$ is only a rough approximation at best and is that only under a response model which assumes that respondents and nonrespondents have the same underlying income distributions. Nonetheless, $Var(\tilde{\tilde{P}}_j)$ is believed to be useful in examining the understatement which exists in the standard CPS variance estimators which ignore the item nonresponse altogether, treating $\tilde{\tilde{P}}_j$ as if (in simple random sampling) it had a variance which could be estimated by

$$\widetilde{Var}(\tilde{\tilde{P}}_j) = \frac{\tilde{\tilde{P}}_j(1-\tilde{\tilde{P}}_j)}{n} \quad .$$

(The proportion $\tilde{\tilde{P}}_j$ is defined to be the average of the two imputations.)

3. RESULTS OF MULTIPLE IMPUTATION APPLICATION WITH THE CPS HOT DECK

This section discusses briefly the impact of the March 1978 CPS income statistics of the hot deck imputation. Four basic tables in Section 7 detail results for persons and families. All data are unweighted and so cannot be directly related to the published information in Consumer Income, P-60, No. 111. In the Proceedings version of this paper we have enough space only to summarize the impact of the imputation on the standard error or CPS income statistics. The conditional standard errors shown for persons and families in figure 2 are the usual variance estimates assuming no missing data (column 4), and the multiple imputation estimates for one, two and 100 imputations (columns 5 to 7). The standard errors for 100 imputations (column 7) approximate the lower limit of the standard error as the number of imputations increases without bound. A measure of the underestimation of the standard error in the usual hot deck variance estimator (column 8) is given by the ratio of column 4 to column 5. Other measures of interest are the ratios of column 5 to column 7 indicating the impact on the variance increase due to performing just a single imputation rather than a great many; finally, the ratio of column 5 to column 6 represents the effect on the variance of imputing twice.

These standard errors on the average (in the first row) confirm a priori speculations as to the direction of the relations among the different variance estimators. In order of magnitude, it appears that--

1. The assumption of no missing data gives rise to the largest effect on the level of the standard error. (We underestimate with this data, conditionally, by 20.0% for persons on the average and 17.0% for families.
2. The failure to adjust for the variance increase arising from a single imputation follows next in importance. (The imputation component of the conditional variance accounts on the average for 13.2% of the standard error for persons and 5.3% for families.)
3. The smallest impact occurs for the multiple imputation itself since it reduces the standard error only slightly when two imputations are conducted (6.1% and 2.6% on the average for persons and families, respectively).

The same pattern noted in figure 2 also appears for each race-sex group separately. There are, however, marked differences in the importance of the extent of the understatement in the standard errors. As figure 3 indicates, the greatest overall impacts occur for males, particularly white males, with the smallest impacts being among females of other races. Other analyses of our multiple imputation experiment can be made with the basic tables in Section 7. We particularly recommend to the reader the information theoretic approach adapted in Report No. 3 of the series, Studies from Interagency Data Linkages (Scheuren, Oh, and Alvey, 1980).

4. SOME CONCLUSIONS AND AREAS FOR FUTURE STUDY

In this paper we have described some of the impacts that imputation for missing data can have on certain uses of the CPS March Supplement statistics published annually by the Bureau of the Census. We have made a particular point of examining the consequences of the fact that the Bureau has historically ignored the variance increases which accompany less than complete reporting of income. Of course, it should be noted that response errors and nonresponse biases are also usually not taken into account explicitly in the CPS (or indeed in most other surveys). The next paper at this session (Oh and Scheuren, 1980b) examines these components of the total error and relates them to the imputation variances developed here.

FIGURE 2.--CPS INCOME SIZE DISTRIBUTION STATISTICS FOR FAMILIES AND PERSONS

(*NUMBERS IN PERCENT*)

INCOME CLASS (IN DOLLARS)	*AVERAGE OF IMPU-TATIONS*	*REPORTED CASES*	*IMPUTED CASES*		*CONDITIONAL STANDARD ERROR*				*STANDARD ERROR UNDER-ESTIMATE*
			FIRST VERSION	*SECOND VERSION*	*UNDER SRS*	*ONE IMP.*	*TWO IMP'S.*	100 *IMP'S.*	
					PERSONAL INCOME				
TOTAL................	100.0000	100.0000	100.0000	100.0000	0.2649	0.3292	0.3104	0.2907	0.8049
UNDER 2,000..........	33.2490	34.9768	25.0380	25.3893	0.1366	0.1703	0.1616	0.1526	0.8017
2,000 *TO* 3,999.......	13.8508	13.6579	14.6743	14.8215	0.1001	0.1143	0.1120	0.1098	0.8763
4,000 *TO* 5,999.......	10.0144	9.8423	11.0473	10.5820	0.0870	0.1387	0.1189	0.0957	0.6275
6,000 *TO* 7,999.......	8.3018	8.2284	8.5596	8.7258	0.0800	0.0949	0.0914	0.0879	0.8427
8,000 *TO* 9,999.......	6.8235	6.7687	7.1591	6.9977	0.0731	0.0876	0.0840	0.0803	0.8348
10,000 *TO* 11,999....	5.8053	5.7265	6.2335	6.1100	0.0678	0.0789	0.0766	0.0743	0.8589
12,000 *TO* 13,999....	4.7715	4.7109	5.0464	5.0608	0.0618	0.0678	0.0677	0.0677	0.9118
14,000 *TO* 15,999....	4.1616	4.0361	4.7949	4.6952	0.0579	0.0665	0.0647	0.0629	0.8706
16,000 *TO* 19,999....	5.3567	5.2029	6.0482	6.0957	0.0653	0.0717	0.0713	0.0710	0.9103
20,000 *TO* 24,999....	3.6235	3.4175	4.4626	4.7000	0.0542	0.0776	0.0685	0.0583	0.6983
25,000 *OR MORE*.......	4.0419	3.4318	6.9360	6.8221	0.0571	0.0632	0.0607	0.0582	0.9034
					FAMILY INCOME				
TOTAL................	100.0000	100.0000	100.0000	100.0000	0.4717	0.5689	0.5547	0.5404	0.8291
UNDER 5,000..........	9.4348	9.9390	7.7051	7.9375	0.1446	0.1826	0.1762	0.1697	0.7920
5,000 *TO* 6,999.......	7.2506	7.5137	6.3089	6.5083	0.1283	0.1603	0.1550	0.1495	0.8002
7,000 *TO* 9,999.......	11.1379	11.3363	10.5079	10.4976	0.1556	0.1797	0.1797	0.1797	0.8661
10,000 *TO* 12,999....	11.2517	11.7507	9.8293	9.4797	0.1563	0.2090	0.1962	0.1828	0.7479
13,000 *TO* 14,999....	7.3747	7.3980	7.2589	7.3411	0.1293	0.1503	0.1493	0.1484	0.8605
15,000 *TO* 17,999....	11.1330	11.4488	10.0658	10.1789	0.1556	0.1835	0.1820	0.1805	0.8483
18,000 *TO* 19,999....	6.5841	6.8037	5.9326	5.8297	0.1227	0.1458	0.1443	0.1428	0.8413
20,000 *TO* 22,499....	7.8667	7.8959	7.8861	7.6599	0.1332	0.1665	0.1598	0.1530	0.8001
22,500 *TO* 24,999....	5.9098	5.8818	5.8811	6.1176	0.1167	0.1501	0.1420	0.1335	0.7771
25,000 *TO* 29,999....	8.8641	8.5191	9.9527	9.9835	0.1406	0.1585	0.1584	0.1582	0.8873
30,000 *TO* 49,999....	10.6839	9.5792	14.2813	14.1579	0.1528	0.1706	0.1687	0.1668	0.8956
50,000 *OR MORE*.......	2.5088	1.9338	4.3903	4.3080	0.0774	0.0817	0.0799	0.0781	0.9476

NOTE: THE TOTAL LINE FOR THE CONDITIONAL STANDARD ERROR COLUMNS CONSISTS OF THE SQUARE ROOT OF THE SUM OF THE STANDARD ERRORS SQUARED FOR EACH CLASS. THE 'STANDARD ERROR UNDERESTIMATE' COLUMN IS THE RATIO IN EACH ROW OF THE ESTIMATED SIMPLE RANDOM SAMPLE (SRS) STANDARD ERROR DIVIDED BY THE ONE 'IMPUTATION' STANDARD ERROR.

FIGURE 3.--CPS STANDARD ERROR UNDERESTIMATES BY RACE AND SEX

(IN PERCENT)

RACE AND SEX	*INCOME STATISTICS FOR*	
	PERSONS	*FAMILIES*
OVERALL............	0.8049	0.8291
WHITE MALES..........	0.7648	0.8058
WHITE FEMALES........	0.8776	0.8419
OTHER MALES..........	0.8787	0.7817
OTHER FEMALES........	0.9218	0.8820

Note: Family data are by the race and sex of the family head. See text for assumptions made. See figure 2 for the income size classes used and the calculation methods employed.

Routine Employment of Multiple Imputation.--There are a number of practical barriers to the routine employment of multiple imputation in a CPS hot deck context. First, we would want to be able to estimate the whole variance including the square of the conditional bias. Second, the complex multi-stage nature of the CPS design needs to be taken into account. In the present paper we were unable to deal with either of these problems; moreover, assumptions about the effect of small respondent cell sizes and other matters had to be made as well.

One simple way to obtain an (over)estimate of the hot deck variance is just to divide the CPS sample into two random halves or pseudo-replicates and do the assignment for missing information separately within each half.

The reason that such a procedure would yield an overestimate is that, using the total sample, one is likely to get a better donor-respondent for each nonrespondent than would be possible with only half the sample. It must be added that we conjecture that any overstatement would be slight given the very large size of the CPS. Furthermore, since there appears to be so little gain in reduced variance from multiple imputations, perhaps the present CPS single imputation estimation procedure should be continued unaltered with the second (half sample) imputations being provided on a separate computer file for use in variance estimation or for other special (or limited) analytic objectives.

5. ACKNOWLEDGEMENTS, AND SELECTED HOT DECK REFERENCES

Acknowledgements.--The work reported on in this paper has benefited greatly from discussions with Gary Shapiro and John Coder at the Census Bureau as well as the discussant at this session, Don Rubin. Barry Fink, at the Census Bureau, and Linda Vogel, at the Social Security Administration (SSA) carried out the second hot deck imputations and created the needed combined computer files. These files, by the way, are available upon request by writing to Henry Ezell at SSA's Office of Research and Statistics, 1875 Connecticut Avenue, Washington, DC 20009. Bob Cripe of the Internal Revenue Service typed the present draft of this paper.

Selected Hot Deck References.--

(1) Bailar, B.A., and Bailar, J.C. III (1979), "Comparison of the Biases of the 'Hot Deck' Imputation Procedures with an 'Equal Weights' Imputation Procedures," Proceedings of the Symposium on Incomplete Data, National Academy of Sciences, Panel on Incomplete Data.

(2) Bailar, B. A. and Bailar, J.C. III (1978), "Comparison of Two Procedures for Imputing Missing Survey Values," American Statistical Association, Proceedings of the Section on Survey Research Methods, 462-467.

(3) Bailar, B. A., Bailey, L., and Corby, C. (1978), "A Comparison of Some Adjustment and Weighting Procedures for Survey Data," Survey Sampling and Measurement, Namboodiri, N. Krishnam (ed.), New York: Academic Press, 175-198.

(4) Chapman, David W. (1976), "A Survey of Nonresponse Imputation Procedures," American Statistical Association, Proceedings of the Social Statistics Section, 245-251.

(5) Colledge, M. J., Johnson, J. H., Pare, R., and Sande, I. G. (1978), "Large Scale Imputation of Survey Data," American Statistical Association, Proceedings of the Section on Survey Research Methods, 431-435.

(6) Coder, J. F. (1978), "Income Data Collection and Processing for the March Income Supplement to the Current Population Survey," Proceedings of the Data Processing Workshop: Survey of Income and Program Participation, U.S. Department of Health, Education and Welfare.

(7) Cox, B. G. (1980), "A Weighted Sequential Hot-Deck Imputation Procedure," American Statistical Association, Proceedings on the Section on Survey Research Methods.

(8) Cox, B. G. and Folsom, R. E. (1978), "An Empirical Investigation of Alternative Item Nonresponse Adjustments," American Statistical Association, Proceedings of the Section on Survey Research Methods, 219-221.

(9) Ernst, L. R. (1980), "Variance of the Mean for Several Hot-Deck Imputation Procedures," American Statistical Association, Proceedings of the Section on Survey Research Methods.

(10) Ernst, L. R. (1978), "Weighting to Adjust for Partial Nonresponse," American Statistical Association, Proceedings of the Social Statistics Section, 468-473.

(11) Fellegi, I. P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Association, 71, 17-35.

(12) Ford, B. L. (1980), "An Overview of Hot Deck Procedures," Incomplete Data in Sample Surveys: The Theory of Current Practices,

National Academy of Sciences, Panel on Incomplete Data.

(13) Ford, B. L. (1976), "Missing Data Procedures: A Comparative Study," American Statistical Association, Proceedings of the Social Statistics Section, 524-529.

(14) Hansen, H., Hurwitz, W. N., and Madow, W. G. (1953), Sample Survey Methods and Theory, Volume II, New York: Wiley, 139-141.

(15) 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.

(16) 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.

(17) Hill, C. J. (1978), "A Report on the Application of a Systematic Method of Automatic Edit and Imputation to the 1976 Canadian Census," American Statistical Association, Proceedings of the Section on Survey Research Methods, 474-479.

(18) Naus, J. I. (1975), Data Quality Control and Editing, New York: Marcel Dekker 107-122.

(19) Nordbotten, S. (1963), "Automatic Editing of Individual Statistical Observations," Conference on European Statisticians, Statistical Studies No. 2, United Nations.

(20) Oh, H. L. and Scheuren, F. J. (1980a), "Estimating the Variance Impact of Missing CPS Income Data," American Statistical Association, Proceedings of the Section on Survey Research Methods.

(21) Oh, H. L. and Scheuren, F. J. (1980b), "Differential Bias Impacts of Alternative Census Bureau Hot-Deck Procedures for Imputing Missing CPS Income Data," American Statistical Association, Proceedings of the Section on Survey Research Methods.

(22) Oh, H. L. and Scheuren, F. J. (1980c), "Weighting Adjustments for Unit Nonresponse," Incomplete Data in Sample Surveys: The Theory of Current Practice, National Academy of Sciences, Panel on Incomplete Data.

(23) Ono, M. and Miller, H. P. (1969), "Income Nonresponses in the Current Population Survey," American Statistical Association, Proceedings of the Social Statistics Section.

(24) Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys -- A Phenomenological Bayesian Approach to Nonresponse," American Statistical Association, Proceedings of the Section on Survey Research Methods.

(25) Rubin, D. B. (1980), "Using Multiple Imputations to Handle Nonresponse," Incomplete Data in Sample Survey: The Theory of Current Practice, National Academy of Sciences, Panel on Incomplete Data.

(26) Sande , G. T. (1979a), "Replacement of a Ten Minute Gap," (with Discussion), Proceedings of the Symposium on Incomplete Data, National Academy of Sciences, Panel on Incomplete Data.

(27) Sande', G. T. (1979b), "Hot Deck Imputation Procedures," Proceedings of the Symposium on Incomplete Data, National Academy of Sciences, Panel on Incomplete Data.

(28) Schaible, W. L. (1979), "Estimation of Finite Population Totals from Incomplete Sample Data: Prediction on Approach," Proceedings of the Symposium of Incomplete Data, National Academy of Sciences, Panel on Incomplete Data.

(29) Scheuren, F. J. (1980), Discussion of Hot Deck papers in the Symposium, Proceedings of the Symposium on Incomplete Data, National Academy of Sciences, Panel on Incomplete Data.

(30) Scheuren, F. J., Oh, H. L., and Alvey, W. L., with Kilss, B. and Del Bene, L. (1980), Matching Administrative and Survey Information: Procedures and Results of the 1963 Pilot Link Study, Social Security Administration.

(31) 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.

(32) Scheuren, F. J. (1976), "Preliminary Notes on the Partially Missing Data Problem -- Some (Very) Elementary Considerations," Methodology Group, Social Security Administration (unpublished working paper).

(33) Scheuren, F. J. (1975), "New March CPS Income Allocation Procedures for Missing Information," unpublished memorandum, Social Security Administration.

(34) Scheuren, F. J. (1970), A Comparison of Selected Economic and Demographic Characteristics from the 1966 and 1967 Surveys of Economic Opportunity and the Comparable Current Population Surveys, Office of Economic Opportunity.

(35) Schieber, S. J. (1978), "A Comparison of Three Alternative Techniques for Allocating Unreported Social Security Income on the Survey of Low Income Aged and Disabled," American Statistical Association, Proceedings of the Section on Survey Research Methods, 212-218.

(36) Spiers, E. F. and Knott, J. J. (1969), "Computer Method to Process Missing Income and Work Experience Information in the Current Population Survey," American Statistical Association, Proceedings of the Social Statistics Section.

(37) 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).

(38) Welniak, E. J., and Coder, F. (1980), "A Measure of the Bias in the March CPS Earnings Imputation System and Results of a Sample Bias Adjustment Procedure," American Statistical Association, Proceedings of the Section on Survey Research Methods.