IMPUTATION OF MISSING ITEMS ON CORPORATE BALANCE SHEETS

Susan M. Hinkins, Internal Revenue Service

The income and financial data in the corporate returns program (Statistics of Income Division, IRS) are compiled from samples of U.S. corporate tax returns, for each tax year. Like most samples and sample surveys, this program must deal with the problems of item nonresponse and data consistency and quality. Aspects of these problems are discussed in two papers presented at these meetings, [1] and [2].

This paper describes one problem of item nonresponse in the corporate returns program. On corporate tax returns, one set of items on each record consists of information from the balance sheet. This paper describes a simulation and analysis designed to investigate the effect of imputation procedures on the final estimates of these balance sheet items. Section I gives a brief background of the problem; Section II describes the experiment and proposed analysis; Section III tabulates the results; and Section IV contains the summary and conclusion.

I. BACKGROUND

The balance sheet is composed of approximately 20 asset items, 10 liability items, and the item Total Assets. The items are defined so that a linear combination, with coefficients 1 or -1, of the asset items must equal the amount in Total Assets, and similarly for the liability items. A list of the balance sheet items is included in Figure 1.

If the balance sheet is missing or incomplete, the missing items are imputed. The current imputation procedure "fills in" the missing items so that they are proportionately consistent with the previous year's totals, by industry and in some cases by the size of total assets.

For national estimates, any effect of the imputation procedure on the estimates is probably negligible. For tax years 1979 and 1978, at most 2% of the (weighted) returns with balance sheets had items imputed on the balance sheet. Even when subdivided by 11 major industrial divisions and 12 asset classes, each cell had a relatively small number of returns with balance sheet items imputed. In 1979 all but one cell had no more than 5% imputed balance sheets and that one cell had 14%. In 1978, all (weighted) cells had fewer than 5%. These tables of nonresponse rates, by industry and asset size, will be published in the complete report, available from the Statistics of Income (SOI) Division.

By taking another level of specific detail, dividing the Financial Division into smaller, though still major, subdivisions, the percentage of returns with imputed balance sheet items becomes appreciably larger - as high as 25%. (We expect this category of returns to have a significant problem with incomplete balance sheets.) Also, the percentage of incomplete balance sheets varies noticeably by major industry and possibly by asset size. The SOI publications provide estimates of balance sheet items to this level of detail.

When the industry subdivision Insurance is further subdivided into 3 categories, Life, Mutual, and Other Insurance, the cross tabulation by asset size produces cells with a large percentage of returns with imputed items, over 50% in several cases. These tables are also available in the complete report. The worst case for tax year 1979 was Mutual Insurance returns with assets less than $100,000; the balance sheets were incomplete on 66% of these returns. We also publish estimates to this level of detail, in the Statistics of Income Source Book for Corporations.

Therefore, although the problem of imputation of balance sheet items may appear insignificant for the purpose of national estimates, it may have a significant effect when smaller subsets of the SOI data file are used. This affects not only SOI estimates and publications; the SOI data base is used by other agencies, for their own research.

Therefore, we have begun a simulation study to investigate the effect of the current imputation method on the estimates, and the possible benefits of an alternative procedure.

II. NATURE OF THE SIMULATION STUDY

There are many factors which may affect the amount and distribution of balance sheet items, the extent of nonresponse, and the effectiveness of an imputation procedure. The factors chosen to control for this experiment are: tax year, major industrial group, asset size, the mechanism for nonresponse, and the percent of returns with incomplete balance sheet. There are (at least) three cases to be considered:

- no balance sheet items are reported,
- Total Assets is reported, and
- Total Assets is not reported but some balance sheet items are reported.

The long range plan is to do a factorial experiment, with these factors, to examine the effect of imputation procedures on the estimates. The ongoing study uses only one combination of these factors.

The Data

The data set is composed of 1979 tax returns which are classified, by industry, as Retail Trade, and which have reported total assets between $50 and $100 million. These returns have complete, edited balance sheets. There are 122 returns in this data set.

The balance sheets on these returns contain 31 items, which are summarized in Figure 1.

The first simulation creates data sets in which the proportion of returns with incomplete balance sheet is 1/3 (40 returns). The degree of incompleteness of the balance sheet will influence how well an imputation procedure works. For simplicity, we start by only considering the case where Total Assets is reported and all the other balance sheet items are missing.

The data sets are generated from the basic

data in the following manner. A subset of 40 returns is chosen randomly. This is the nonresponse mechanism; it generates data that are missing at random. Call this set of records A and the remaining (82) returns C. On the returns in set A, delete all the items on the balance sheet except Total Assets. These "missing" items are then imputed using each of two procedures described in the next section, yielding 2 sets of records, A_1, and A_2. Two sets of estimates, based on the following records C and A_1,
C and A_2,
can then be compared to the original, correct totals, using C and A.

This simulation and estimation procedure was replicated 10 times to produce estimates of bias and error due to imputation, under these conditions. A replicate is generated by randomly selecting a new set of 40 records.

Two Imputation Procedures

The first procedure tested will be the method currently used to impute balance sheet items on corporate returns. When only Total Assets is observed, the current imputation procedure imputes missing items in the following manner. One set of (4) items is always imputed as 0. These items will be underestimated, resulting in biased estimates, and the shape of the frequency distribution may be substantially altered. [13] Another set of items is imputed to 0 only if related income items are not present (0) on the return. For example, it there is no Depreciation Deduction reported, the balance sheet item Depreciable Assets is imputed to 0. If the income item is present, the related balance sheet item is imputed in the same way as the third class of items. For the third (and last) class of items, the procedure basically allocates the amount to be imputed by dividing up the unreported amount of Total Assets in the same relative proportions as the previous year's estimates of these items. Unless the relative proportions are the same this year as last, this procedure results in biased estimates.

The second method is a variation on the present method; a hot deck error term is added. Hot deck procedures have been used on other survey data bases, and their advantages and disadvantages discussed in the literature [8]. However the term has been used with several definitions. The procedure we used is as follows. Let a be a record (return) from the set A and let the i^{th} balance sheet item to be imputed be denoted by $x_i(a)$. Using method 1 (the current procedure), an estimate $\tilde{x}_i(a)$ is calculated. Select a return, call it c, with 1) the same minor industry classification as a, and 2) a complete (observed) balance sheet. The selection is done in the "hot deck" manner, i.e. select the last such return that was seen, as in a card deck. Because the balance sheet is complete on c, the i^{th} item, $x_i(c)$, is known. Delete the balance sheet items except for Total Assets from return c and apply the current imputation procedure to get estimates $\tilde{x}_i(c)$. Then the hot deck error term for each i is calculated as $e_i = x_i(c)-\tilde{x}_i(c)$. The estimate of $x_i(a)$ using this procedure is

$$\tilde{x}_i(a)+e_i = \tilde{x}_i(a)+x_i(c)-\tilde{x}_i(c).$$

This procedure is an attempt to use a more phenomenological approach; that is an attempt to make the imputed values more similar to observable values. [9] This might be an improvement for users of our data who look at microdata, small subsets or individual records. It also uses the most current observed values for estimating the unobserved. If the nonresponse mechanism is ignorable, as in this simulation, this procedure results in unbiased estimates, where the expected value is taken over repeated random samples of the same size and with the same degree of nonresponse. It may also give better estimates of variance.

III. SIMULATION RESULTS

Imputation is one method of modeling the unobserved values based on the observed. The practical motivation for using imputation procedures to handle nonresponse is that it allows the use of standard, complete data techniques and analysis.

One of the chief difficulties, of course, is that the nonresponse mechanism may not be ignorable and the unobserved may have a different distribution than the observed responses. In this case, the nonresponse mechanism should be modeled and used in the imputation procedure. The nonresponse mechanism was certainly not random across all returns (see Section I). However it is possible that the mechanism may be reasonably modeled as random within classes defined by controlling certain background variables, such as industry classification and size of assets. Therefore, the use of an ignorable nonresponse mechanism in the simulation is not unrealistic.

If the data are missing at random, as in this simulation, a valid inference for the mean is the usual sample mean and standard error using only the observed records (82 records in our study). Such a procedure for the Statistics of Income data base would be very complex and impractical; each item would need its own weighting factor, to reflect the number of records on which it is observed. By imputing missing items, the sample estimates can be calculated across all records, independent of the degree of nonresponse. For example, in this study, the sample mean is calculated across all 122 records, including the 40 records with imputed items. This ease of analysis is especially important in complex, multi-purpose surveys such as the Statistics of Income data. However the resulting estimates and inference should be comparable to the results using complete data only, in the case of data that are missing at random.

Therefore the results using imputed data are compared to those calculated using observed data only. For each item, three sample values are compared in this way for each imputation procedure:

1) The average relative bias of the sample mean (%), $\frac{\text{average bias of X}}{\text{true value, X}}$. (100%).
2) The percentage of estimated 95% confidence intervals for X that included the true value. (There is a confidence interval calculated for each replication, for a total of 10.)
3) The average width of the confidence intervals.

Excluding for the moment those items that are always imputed to 0, the current procedure is similar to a best prediction method for each return. It models the unobserved by using the estimated totals from the previous year. If the relationships are the same this year as last, it should result in unbiased estimates of the means. If not, the resulting estimates will be biased. Because the simulation uses an ignorable nonresponse mechanism, adding the hot deck error term should result in unbiased estimates. Neither procedure would necessarily correct for bias if the distribution of the values for nonrespondents were significantly different than the respondents'.

Based on analytical results in the current literature (for example, Rubin, [10]) we expect the confidence intervals for the sample mean using the current procedure to be too narrow. Adding the hot deck error should result in wider intervals, though probably still not wide enough. The current imputation procedure (CIP) underestimates the variability more than the hot deck procedure because the CIP tends to "stack" the imputed values in the same location. Neither procedure correctly estimates the variability because, as Rubin [10] points out, they provide no estimate of the additional variability due to the imputation procedure - the distribution used to generate the imputed values. Rubin recommends the use of multiple imputation to allow calculation of valid standard errors of the estimates.

By adding the hot deck term, we hope to impute records that better represent observable values, that maintain the original distribution. Therefore, for two items, we also compared the resulting frequency distributions when the imputation procedures were used, to the original distribution.

For the returns in this simulation study, there are four items that are always imputed to 0 under the current imputation procedure. This will result in:

- biased estimates of the means,
- incorrect inference, 95% confidence intervals that do not contain the true value with probability .95,
- significantly altered frequency distributions compared to the true distribution.

We expect the proposed imputation procedure, including the hot deck error term, to provide much more acceptable data.

Figure 2 shows the resulting average relative bias for each of these four items. As expected, using the CIP the mean is vastly underestimated, by as much as 50%. The second imputation procedure, including the hot deck error term, (HDP), significantly reduced the bias, and in this respect did as well or better than estimation based on observed values only. Figure 3 tabulates the percent of estimated 95% confidence intervals that contained the true value, based on the 10 replicates. As expected, the coverage properties using the CIP was unsatisfactory; the coverage was greatly improved using HDP. Figure 4 shows the ratio of the average width of the 95% confidence interval based on data including imputed records, to the average width of the interval based on observed data only. As expected, the intervals are too narrow using either imputation procedure, but those associated with the HDP are generally the wider of the two.

The remaining balance sheet items are imputed by allocating the amount to be imputed in the same relative proportions as the results from the previous year. There are nine items that are imputed in this way only if related income items are present, otherwise they are set to 0. The CIP may result in biased or unbiased estimates, depending on the relative stability of the ratios from year to year. The HDP should result in unbiased estimates.

Figure 5 shows the relative bias for these items, i.e. the items that are always imputed and those that are only imputed if related income items are present. For 29 of these 30 items, the HDP resulted in estimates with smaller or approximately equal relative bias compared to the CIP. In that one case, the HDP did as well as the estimate using only the observed values, but neither did as well as the CIP. This was also the only item in which the relative bias using the HDP was greater than 15%.

For most of these items, the CIP generally resulted in unbiased estimates. So a reduction in bias using the HDP is not of major interest. However there were 5 items (Investment in Government Obligations (U.S.), Depletable Assets, Accumulated Depletion, Amortization, and Mortgages (Less than 1 year)) for which the CIP resulted in estimates with a relative percent bias over 15%. In 4 of the 5, the mean was underestimated. Presumably this bias is at least partially due to the use of the previous year's ratios when they are no longer an adequate prediction of the present year's relationships. For example consider the item Depletable Assets which had the largest relative basis (-37%). This item was underestimated, and in fact the ratios (based on 1978 data) used to estimate the 1979 data were generally smaller than the final ratios calculated from the 1979 file.

For all 5 items, the HDP resulted in an appreciable reduction in bias.

The tabulation of coverage properties of the estimated confidence intervals can be found in the complete report, available from SOI. The two procedures were generally comparable in coverage properties. The HDP showed a significant improvement for only one of the items, an item that was overestimated using the CIP.

In general, the width to the confidence intervals using the CIP will be too short compared to the intervals estimated from the observed data only. The intervals when the HDP is used should be wider that those using the CIP, but they will generally still be too short. This effect was also demonstrated; the data are available in the complete report.

From the analysis so far, we can conclude that for inference about means, the CIP resulted in adequate estimates and inference for most items, but not for all. The HDP provided better estimates of the mean (unbiased) and the correct inference for all the items.

In addition, the Statistics of Income data are used by various clients, for other purposes. The shape of the distribution of specific items on certain classes of returns may often be of interest, i.e. the microdata. For such purposes the CIP results in data that are less satisfactory.

The problem is obvious for the variables currently imputed to 0. This procedure will alter the distribution and may create records that do not resemble observable records. Figure 6 shows this effect on one such variable, Beginning Inventories, based on the first replicate. The distribution of Beginning Inventories as a fraction of Total Assets is shown for those corporations in this study that were classified as Food Stores. There were 28 such records, of which 8 were selected for imputation. The first histogram shows the original distribution - all 28 observed values. The second graph shows the resulting distribution when 8 of the 28 records were imputed using the CIP (i.e. imputed to 0). The last histogram shows the resulting distribution when the HDP was used for imputation. The results are as expected. The resulting distribution using the CIP is much different that the true distribution. Using the HDP we do not recreate the original distribution exactly, but the correct shape and general properties of the frequency distribution are maintained.

Next consider an item that is not necessarily imputed to zero, a variable for which the CIP did as well as the HDP in the previous analysis of inference about the mean. Consider the item Fixed Depreciable Assets. The relative bias using either the CIP or the HDP was almost 0 and the coverage by the estimated confidence intervals was 100%. Again consider Fixed Depreciable Assets as a proportion of Total Assets for the 28 records classified as Food Stores. Figure 6 also shows the original (true) distribution, the distribution resulting when 8 of the 28 records are imputed using the CIP, and the resulting distribution when the HDP is used. This demonstrates the effect of the CIP in altering the distribution by "stacking" all imputed values at approximately the same point. This results in a much sharper distribution - with smaller variance. The HDP results, in this case, in a distribution that is somewhat too heavy-tailed, with increased variance. But the HDP results in a frequency distribution that is much more similar to the original distribution.

IV. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE STUDY

The hot deck imputation procedure generally improves upon the current imputation in that it produces:

- unbiased estimates of the means,
- better estimates of the standard error,
- somewhat better coverage by the estimated confidence intervals,
- improved microdata, imputed values that better represent the original frequency distribution.

Improvements using the hot deck error term are especially dramatic for those items that are currently imputed to 0. Imputing the value 0 into all missing records has only one redeeming feature; it is easy. However it results in biased estimates, incorrect inference, and records that do not resemble observable values.

There are several issues associated with imputation that have not been addressed in this discussion. The recent article by Sande [11] provides a good description of the trade-offs between imputation strategies and some other general issues that need to be addressed for our specific problem. The relationship between editing and imputation requires further consideration. The hot deck procedure requires additional consistency tests; the imputed data must satisfy the edits, must be consistent. For the relatively simple case simulated in this study, this problem was easy to solve. For other classes of returns, with other patterns of missingness, this problem may be more complex.

A problem specific to a hot deck procedure is the possibility of having too few "similar" returns with complete, observed balance sheets. A strategy must be developed to balance the effect of reusing the observed records with the effect of relaxing the definition of "similar" returns.

The simulation study assumed data that were missing at random. The returns with incomplete balance sheets are clearly not randomly distributed across asset size and industrial classification. However the assumption of missing at random may be valid within classes defined by asset size and industry. However, if this is not the case, neither imputation procedure will result in representative microdata or correct inferences.

For some imputation problems, we can safely assume an ignorable nonresponse mechanism, because we generate the nonresponse. In order to cut processing costs, in the future we would like to designate that certain items will not be retrieved from each return. An example of this is given in the paper presented at these meetings, by Cys, Hinkins, and Rehula [2].

Imputation procedures should be investigated using additional information where available. For example, observed balance sheet information for a particular return may be available from the previous year.

In conclusion, while there are still issues to be considered and problems to be solved, there is evidence that adding a hot deck component to the imputation procedure will result in significant improvements in our data. Further work in these areas is being pursued.

Figure 1. Mean Value for the Balance Sheet Items Reported on the (122) Returns in the Simulation

Item Number	Title	Amount in $1000
33	Beginning Inventories	17,973
34	Cash	4,165
35	Trade Notes and Accounts Receivable	8,796
36	Allowance for Bad Debts	342
37	Ending Inventories	20,407
38	Investment in Government Obligations - U.S.	576
39	Investment in Government Obligations - States	7
40	Other Current Assets	3,725
41	Loans to Stockholders	358
42	Mortgage and Real Estate Loans	377
43	Other Investments	5,100
44	Fixed Depreciable Assets	34,000
45	Accumulated Depreciation	11,936
46	Depletable Assets	103
47	Accumulated Depletion	6
48	Land	2,865
49	Intangible Assets	690
50	Accumulated Amortization	151
51	Other Assets	2,334
52	Total Assets	71,067
53	Accounts Payable	11,929
54	Mortgages, Less than 1 Year	3,914
55	Other Current Liabilities	7,287
56	Loans from Stockholders	281
57	Mortgages, 1 Year or More	15,549
58	Other Liabilities	2,523
59	Capital Stock, Total	4,024
60	Paid-in or Capital Surplus	5,755
61	Retained Earnings - Appropriated	519
62	Retained Earnings - Unappropriated	19,878
63	Cost of Treasury Stock	592

Figure 2. Average Relative Bias of the Mean (Items Imputed to Zero)

Item Number	Percent Bias		
	Current Procedure	Hot Deck Procedure	Observed Values Only
33	-33	1	- 7
41	-34	4	- 4
56	-49	-17	-24
63	-28	7	- 4

Figure 3. Percent of 95% Confidence Intervals that Contained the True Mean.

Item Number	Estimation Based on:		
	Current Procedure	Hot Deck Procedure	Observed Values Only
33	0	100	80
41	80	90	90
56	40	70	90
63	70	100	80

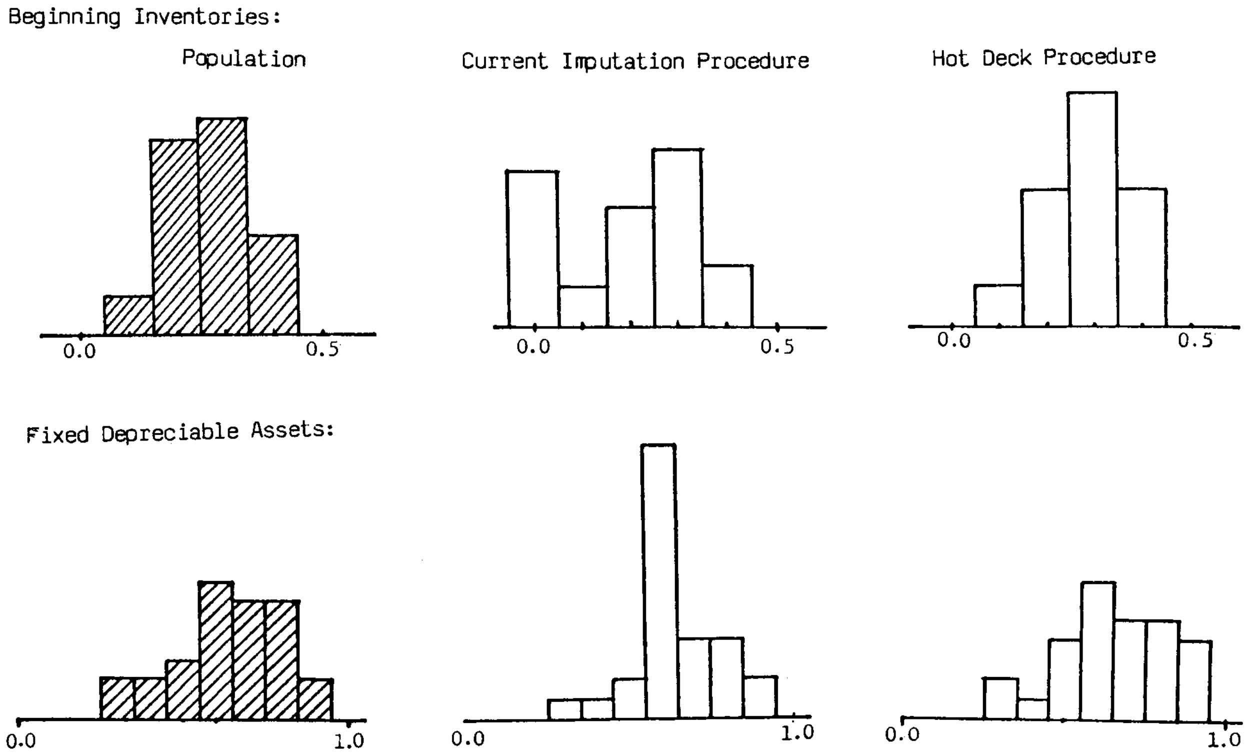
Figure 4. The Ratio of the Average Width of the 95% Confidence Intervals Using Imputed Data, to the Width of the Interval Using Observed Data Only.

Item Number	Current Procedure	Hot Deck Procedure
33	.79	.73
41	.79	.96
56	.75	.96
63	.72	.85

Figure 5. Average Relative Bias of the Mean

Item Number	Percent Bias		
	Current Procedure	Hot Deck Procedure	Observed Values Only
Part I: Asset Items that are Always Imputed			
34	6	2	2
35	10	2	- 1
36	- 6	- 2	- 3
37	6	1	0
40	-10	2	- 1
42	- 4	0	- 5
48	- 6	1	1
49	- 2	- 5	-18
50	17	7	-12
51	- 6	- 5	1
53	2	- 1	- 2
55	0	4	2
58	- 8	- 6	3
59	12	- 3	- 1
60	-10	5	2
61	-12	13	- 4
62	- 5	0	0
Part II: Items that are Not Always Imputed			
38	-27	- 5	9
39	- 4	20	21
43	- 3	- 2	- 1
44	- 3	0	- 1
45	2	2	0
46	-37	- 7	- 6
47	-35	- 4	- 6
54	30	4	2
57	- 1	- 2	- 3

Figure 6. Illustration of Frequency Distributions for Two Balance Sheet Items



ACKNOWLEDGMENTS

The author would like to thank numerous members of the Corporation Returns Analysis Section for their contributions. Appreciation is also extended to Wendy Alvey and Beth Kilss for editorial help with the manuscript, and to Sandra Heaney for typing numerous drafts.

NOTES AND REFERENCES

[1] Bahnke, J.E. and Wheeler, T.D., "Corporate Statistics of Income: Data Testing," 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[2] Cys, K., Hinkins, S. and Rehula, V., "Automatic and Manual Edits for Corporation Income Tax Returns," 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[3] Greenberg, B., "Using an Edit System to Develop Editing Specifications," 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[4] Greenless, J.S., Reece, W.S. and Zeischang, K.O. (1982), "Imputation of Missing Values When the Probability of Response Depends upon the Variable Being Imputed," Journal of the American Statistical Association, Vol. 77, pp. 251-261.

[5] Kalton, G. and Kasprzyk, D., "Imputing for Missing Survey Responses," 1982 American Statistical Association Proceedings Section on Survey Research Methods.

[6] Little, R.J. (1982), "Models for Nonresponse in Sample Surveys," Journal of the American Statistical Association, Vol. 77, pp. 237-250.

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

[8] Oh, H.L. Scheuren, F.J. and Nisselson, H., "Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data," 1980 American Statistical Association Proceedings, Section on Survey Research Methods.

[9] Rubin, D.B., "Multiple Imputations in Sample Surveys -- A Phenomenological Bayesian Approach to Nonresponse," 1978 American Statistical Association Proceedings, Section on Survey Research Methods.

[10] Rubin, D.B. (1980), "Handling Nonresponse in Sample Surveys by Multiple Imputations," Monograph prepared for the Census Bureau.

[11] Sande, I.G. (1982), "Imputation in Surveys: Coping with Reality," American Statistician, Vol. 36, No. 3, pp. 145-152.

[12] Internal Revenue Service Source Book, Statistics of Income--1979 Corporation Income Tax Returns, Statistics of Income Division, 1982.

[13] These items are listed in Figure 2. The documentation of the current imputation procedure does not explain why these items are always imputed to zero.