AN EXAMINATION OF IMPUTATION FOR ASSET ITEMS IN THE NEW BENEFICIARY SURVEY

Robert L. Santos and Carlos G. Lazaro
Institute for Survey Research, Temple University
Philadelphia, PA 19122

1. Introduction

Imputation is the process of replacing missing survey data by pseudo-values or estimates; this type of nonresponse adjustment yields a "completed" data set with no apparent missing items. Many empirical investigations of alternative imputation procedures exist in the literature. (See Ford 1976, Bailar and Bailar 1978, Scheiber 1978 and David et. al. 1986, to name a few.) In this paper, however, we take a different approach.

We select seven assets from the 1982 New Beneficiary Survey (NBS) for which missing data were imputed. We assess the impact of imputation on amounts and income flows from these assets. By conducting an "after-the-fact" assessment, we illustrate a long neglected component of imputation methodology: Examination of imputations to uncover effects on the univariate structure of the data.

This paper has three principal objectives:

(i) to assess how imputation affects the data set;
(ii) to assess the plausibility of imputations; and,
(iii) to make recommendations to analysts of NBS data in light of the findings in (i) and (ii).

Seven assets are selected for analysis: money market, checking, savings and credit union accounts, certificates of deposits, bonds and stocks. Two asset variable types are analyzed: asset amounts (the value of asset) and asset income flow (the interest or dividend return). We examine aggregate net worth of assets, total net and individual return rates from assets, and aggregate net income flow from assets.

To see how imputation affects asset data, totals and averages from the respondent data are contrasted with those of imputed data. Then we compare respondents to the total combined sample.

Two nonresponse models are employed to assess the plausibility of imputation. The first decomposes the sample according to 19 patterns of possessed assets. The second partitions the sample into 30 subgroups reflecting combinations of three significant predictors used in all imputation schemes. The nonresponse models postulate that responders and nonresponders have similar characteristics within subgroups. Under this assumption, the plausibility of imputation is assessed by simply comparing respondents to nonrespondents within subgroups.

The true values of the nonresponders can never be known. Thus, the limitations of our investigation must be recognized. An assessment of imputation plausibility is ultimately subjective because it requires the postulation of a nonresponse model. We believe the models employed in our analyses are reasonable. However, we concede the possibility that other nonresponse models could yield different conclusions.

As a final introductory comment, we note that imputations cannot be expected to withstand an exhaustive, detailed scrutiny of their appropriateness for various analyses. Imputation is a general nonresponse compensation technique intended to be adequate, not optimal for a variety of analyses. In reviewing the imputation methodology we generally believe the techniques employed were reasonable. However, it is important to uncover some of the strengths and weaknesses of the resultant imputed data set. This is our intention.

2. The New Beneficiary Survey

The 1982 New Beneficiary Survey (NBS) was conducted by the Institute for Survey Research under contract from the Social Security Administration. The survey utilized a national sample of first time recipients of Social Security benefits to examine employment history, income and asset holdings. The data were gathered to examine policy issues relating to Social Security program changes.

The sample represents four specific domains of study: Retired Workers, Disabled Workers, Spouses (wives and widows, including those who were divorced), and a Medicare-only group (i.e., medicare beneficiaries who are eligible for Social Security but do not receive it). A clustered multi-stage probability sample was employed. A total of 18,599 interviews were conducted with an overall response rate of 86 percent. Complete documentation of the NBS methodology is furnished in The 1982 New Beneficiary Survey: Users Manual (1986).

3. Missing Data and Imputation in the NBS

All missing asset items were imputed using stochastic techniques. In the next section, we present missing data rates for each asset variable and sketch the general imputation process. A detailed analysis of missing data in the NBS is presented in Mattlin (1987). A full documentation of imputation methodology is found in Czajka (1984), and a summary is provided in the NBS Users Manual.

3.1 Asset Amounts

Column 2 of Table 1 presents the nonresponse rates for seven asset amounts. Missing data rates ranged from about one in six for checking accounts to over one in three for stocks. Thus, imputation was substantial.

The imputation process began by predicting the total net worth from all assets exclusive of home equity. It was then disaggregated among assets held. The prediction equation was estimated by regressing the log net worth of responders on a large set of predictors. Predicted net worths were summed with stochastic terms generated from the respondent empirical distribution of residuals. The result constituted imputed net worth.

To disaggregate net worth, predicted shares across assets were calculated from respondent data and applied to nonresponders. If all amounts were missing, imputed net worth was

apportioned in proportion to predicted shares. Adjustments were made when partial asset amounts were reported.

Table 1. Missing Data Rates among Asset Amounts and Income Flows in NBS

Asset Items	No. of Reported Holders	Asset Amounts % Missing	Income Flows % Missing
Money Market	3,965	26.6%	37.7%
CD	5,100	25.6	38.6
Savings	10,320	21.1	37.5
Credit Union	2,398	18.6	37.5
Checking	13,584	16.3	21.5
Bonds	2,591	32.2	64.5
Stocks	2,804	36.2	31.6

3.2 Income Flows from Assets

Rates of missing data for asset income flows are reported in Column 3 of Table 1. Nonresponse is highest among flows, ranging about one in five cases for checking accounts to almost two in three for bonds. Nonresponse rates exceed 30 percent for all but one asset. Consequently, imputation was substantial.

Asset income flows were imputed stochastically. A zero versus positive income indicator was first imputed. Imputations were then calculated among those with positive income flows. To determine positive/zero income flags, proportions of respondents without income from assets were calculated for ranges of asset amounts by asset type. Income flags were determined by comparing a randomly generated probability to the observed proportion.

To impute positive income flows, the expected log rate of return was calculated as the sum of a mean log rate and the product of a random standard normal deviate and a standard deviation. The imputed income flow comprised the product of the reported asset amount and the exponentiation of the expected log rate of return.

4. Analysis of Asset Amounts

4.1 Net Worth of Seven Assets

Net worth of assets was imputed to nonresponders via stochastic regression. Imputed net worths were then disaggregated into individual amounts using a proportional allocation scheme. Because of the two step nature of the imputation process, we begin with an examination of total net worth from seven assets.

It would be inappropriate to compare the average net worths between full reporters and imputed cases for two reasons. First, imputation tended to overstate the number of assets in possession relative to full reporters. (See Santos and Lazaro 1987 for details.) Consequently, total imputed net worths would likely be larger than those reported. Secondly, by their very nature, net worths of assets can fluctuate from person to person. To control the effects of these factors, we decomposed the sample into 19 asset holdings patterns shown in Table 2. Average net worths of reporters were then compared to those imputed.

Column 2 depicts this comparison by presenting the ratios of imputed to reported mean net worth. Two thirds of the cell means of imputed cases are larger than those of reporters. Average imputed net worths for these patterns are one third to over twice as large as their reported counterparts. The 6 remaining imputed cell means are 12 to 85 percent smaller than the corresponding reported means. However, these cells account for less than 10 percent of the total sample, and thus may be considered negligible. We conclude that imputed net worth displays a modest positive bias under this nonresponse model.

Column 3 exhibits the ratios of total sample to reported sample means. This shows the effect of imputation on the total combined sample. Average net worths exceed those reported by 6 to 58 percent in 11 of 19 asset holdings patterns. These patterns account for 85 percent of all cases. Six cell means are 5 to 30 percent smaller than the corresponding reported means, but collectively these patterns account for less than 10 percent of the sample. Under this nonresponse model, imputation creates an average positive bias ranging from 6 to 58 percent.

4.2 Amounts of Individual Assets

The investigation of the effects of imputation on individual asset values is conducted in two parts. First, we consider the contribution of imputed data to estimated totals. Secondly, we examine the effect of imputation on average amounts of assets.

Estimated totals can be expressed as $Y = Y(r) + Y(m)$, where $Y(r)$ represents the total from the reporters and $Y(m)$ denotes the imputed total. Total population dollar values were estimated for each asset and decomposed into contributions from reported and imputed cases. Depending on the asset, imputed cases accounted for 32 to 41 percent of the estimated total population dollar value. Such large contributions are alarming. Irrespective of imputation plausibility, these estimates of population totals should be cautiously interpreted. It would be prudent to avoid estimation of total asset amounts altogether, or perhaps use more sophisticated methods of estimation which are tailored to handle missing data.

Next we considered the effect of imputation on estimates of mean asset amount. For each asset, we calculated mean asset amounts among reporting cases and imputed cases. Apart from stocks and bonds, the imputed average asset amounts were 33 to 69 percent larger than their reported counterparts. The average imputed face value of bonds was 11 percent under the reported average, while the average imputed value of stocks was 12 percent larger.

To assess the effect of imputation on the total sample, ratios of overall to reported average amounts were calculated for each asset amount. Apart from stocks and bonds, overall average asset amounts exceeded those reported by roughly 10 to 15 percent. Average asset values of imputed stocks and bonds were respectively larger and smaller than their reported averages

by 4 percent. Thus, imputation of asset amounts had a substantial impact on the survey data.

To examine the question of plausibility of imputations, we created a second nonresponse model. This model decomposes the sample into 30 subpopulations or "cells." These cells represent a cross-classification of three major predictors employed in the stochastic imputation methodology. The three variates collectively represent significant predictors of asset amounts and flows. They are:

(i) Demographic Group - a fivefold categorization combining marital status, retirement status, sex and disability status.

(ii) Education - a threefold categorization of completed years.

(iii) Primary Insurance Amount - a variable from SSA's Master Beneficiary Record which reflects the social security benefit level. PIA was dichotomized into equal halves using the median as a breakpoint.

The plausibility of imputation was checked by comparing reported and imputed average amounts within cells. Generally we expect reported and imputed mean amounts to be similar within cells, since we are conditioning on the variates employed in the imputation process.

Rather than presenting a thirty-by-two table of averages, we graphically compared averages via scattergrams. Figure 1 plots respondent vs. imputed mean asset amounts within cells; the scattergram overlays the results for seven asset types. This was done for the sake of parsimony, since the individual scattergrams exhibited the same tendencies.

Under the 30 cell "nonresponse model", imputations are "plausible" on average, if the scatter of reported vs. imputed rates follows the Y=X line through the origin. Figure 1 shows that the plotted means favor a general linear relationship, but that substantial variation exists. Moreover, most imputed mean amounts exceed reported cell means for all asset types. This is evidenced by the larger number of observations above the Y=X line. We therefore conclude that, on average, imputed asset amounts tend to be positively biased under this nonresponse model. Moreover, much of the bias will be retained in the overall sample, since rates of imputation ranged from 16 to 36 percent among asset types.

5. Analysis of Income Flows from Amounts

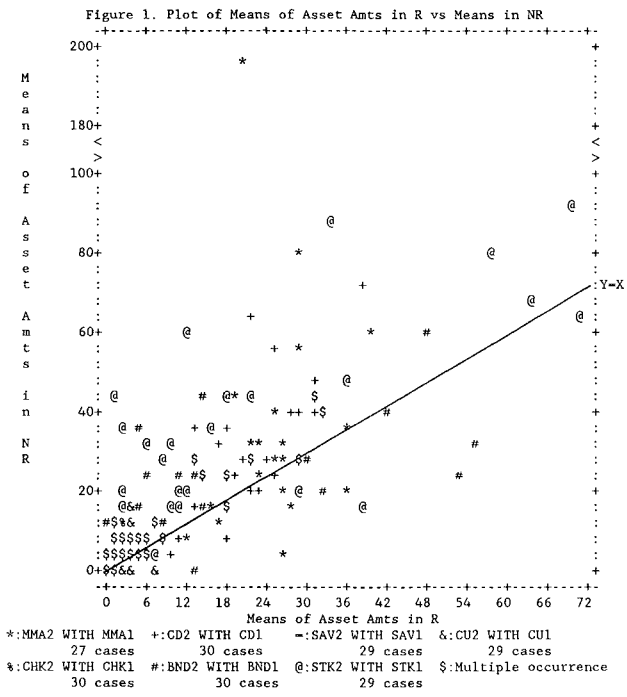
Income flow from assets denotes the interest or dividend accumulated over a specified time period. In consequence, income flow involves two distinct items: an amount (of income flow), and a time period (corresponding to the flow). In the analyses that follow, income flow denotes the annualized return from an asset. An imputed flow has had the flow amount, the time period covering the amount, or both imputed.

Asset income flows were imputed by applying rates of return to amounts of asset held. Therefore, two items are examined. The first is the percentage rate of return, calculated from the ratio of income flow to amount held. The second is simply the actual dollar amount of income flow.

Table 2. A Comparison of Reported, Imputed and Total Averages for Net Worths, Return Rates and Total Income Flow from Seven Assets in the NBS

Pat. No.	Holding Pattern* Ck	Sv	CD/MM CU	S/B	Average Net Worth Fully Reported ($'000) (1)	Imp./ Rep. Ratio (2)	Tot./ Rep. Ratio (3)	Average Return Rates Fully Reported (%) (4)	Imp./ Rep. Ratio (5)	Tot./ Rep. Ratio (6)	Ave. Total Inc. Flow Fully Reported ($'000) (7)	Imp./ Rep. Ratio (8)	Imp./ Rep. Ratio (9)
1	-	-	-	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	H	H	-	-	5.3	172%	119%	8.2	60%	83%	0.35	111%	106%
3	H	H	1	-	27.0	137	111	11.1	81	91	3.00	100	100
4	H	-	-	-	0.7	214	114	3.1	48	90	0.02	200	116
5		H	1	1-2	76.7	135	116	10.2	83	88	7.92	91	94
6	H	-	1	-	20.8	131	106	9.7	88	96	2.47	89	96
7	H	H	2-3	1-2	115.2	137	118	9.7	99	99	12.95	95	96
8	-	H	-	-	2.7	104	100	11.5	57	83	0.22	64	86
9	H	H	2-3	-	42.1	140	119	10.6	92	95	5.04	105	103
10	H	H	-	1-2	35.4	73	87	7.9	84	89	2.90	62	72
11	H	-	1	1-2	101.1	88	95	11.4	98	99	10.32	72	83
12	H	-	2-3	-	37.9	153	118	10.2	101	100	4.54	100	100
13	-	H	1	-	15.8	244	147	12.5	68	82	1.88	136	120
14	H	-	2-3	1-2	98.2	156	124	9.1	103	102	11.68	104	103
15	H	-	-	1-2	105.3	15	70	5.4	1,633	1,011	4.36	27	57
16	-	H	0-1	1-2	126.6	38	74	10.8	79	84	3.58	108	106
17	-	-	1-3	-	11.8	22	91	11.3	1,927	696	1.34	50	84
18	-	H	2-3	0-2	56.1	202	158	10.2	89	92	7.01	120	114
19	-	-	0-3	1-2	38.3	21	79	19.8	56	75	3.51	39	65

* An 'H' signifies that the asset is held; a dash (-) denotes nonholding status. The two other columns include the following: money market, certificates of deposit, credit union, stocks and bonds. Under this columns, a zero (0) represents nonholding status to the corresponding assets; number ranges represent the number of assets which may be held (e.g., 1-2 denotes at least one and up to two of the assets held).



Imputed income flows could be biased for three reasons: (i) the asset amount could be biased; (ii) the rate of return could be biased; or (iii) both could be biased. Thus, analysis of both return rates and flow amounts yields a more meaningful assessment of the effects of imputation.

5.1 Analysis of Net Rates of Return

We consider the overall net rate of return because analysts are concerned with total income, including income from assets. As such, the income flow and rates of return may prove to be plausibly imputed at the aggregate level, but not at the individual asset level.

Column 4 of Table 2 presents the average percentage return rates for responders across 18 specific patterns of asset possession. A net rate of return was deemed "imputed" if at least one of seven flows was imputed. Column 5 presents the percentage ratios of imputed to reported average rates of return. Patterns 2 through 14 contain 95 percent of all cases with asset holdings. Restricting attention to these patterns, all but two average imputed rates of return are smaller than their reported counterparts. In 8 of these patterns, average imputed return rates are 12 to 52 percent lower than reported rates.

Column 6 exhibits the ratios of total sample to respondent average return rates. In six of the first thirteen patterns, overall average return rates are within 5 percent of reported rates. In the other seven patterns, overall average return rates are within 10 to 17 percent of reported rates.

The ratios in Columns 5 and 6 contrast sharply with those of Columns 2 and 3. The average imputed net worth was positively biased, while average imputed return rates are slightly negatively biased. Of course, this assumes that the 19 asset patterns account for differences between responders and nonresponders.

5.2 Individual Rates of Return

Next, we consider the rates of return for individual assets. For each asset, average imputed and average respondent return rates were calculated. We found that return rates were highest among stocks, money market accounts and certificates of deposit; they were lowest for checking accounts. On average, imputed return rates were almost twice the reported rates for stocks, and roughly one fifth larger than reported rates for money market accounts. For the remaining assets, imputed return rates were 7 to 50 percent lower than reported rates. The largest disparities occurred in savings, credit union and checking accounts, where imputed average return rates were one third to one half smaller than reported rates.

To see the effect of imputation on the total sample, ratios of overall to respondent average return rates were calculated for each asset. Overall return rates of stocks and money market accounts were 44 and 8 percent larger than reported rates, respectively. Overall return rates of CDs and bonds were only slightly less than those reported. For savings, credit union and checking accounts, the overall return rates were about one eighth to one sixth smaller than the corresponding reported rates. Thus, the effect of imputation on average return rates varied substantially from asset to asset.

To check the plausibility of imputed return rates, the sample was decomposed according to our 30 cell nonresponse model. Figure 2 plots reported cell mean rates of return against those imputed, and overlays the plots for all seven assets. Plotted points tend to fall below the Y=X line, suggesting that return rates are slightly negatively biased under our nonresponse model. However, the behavior of individual asset return rates is not as consistent as those of Figure 1. Stocks and bonds tend to have roughly equal numbers of points above and below the Y=X line, suggesting no systematic or slight negative bias for imputed return rates. Money market accounts, which over the entire sample have higher imputed average return rates than those reported, show more points below the Y=X line, indicating a stronger negative bias.

5.3 Analysis of Net Flows from Assets

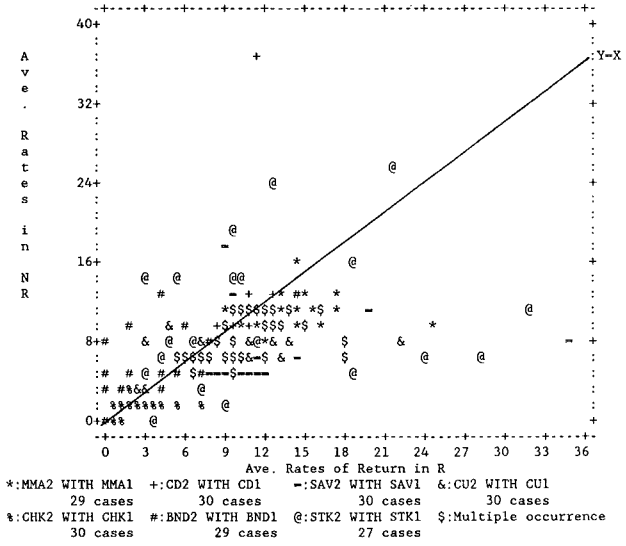
We now consider the aggregate income flow from all seven assets. From Section 5.1 we concluded that, in general, imputed net return rates were slightly biased below the reported rates. Also, in Section 4.1, a modest positive bias for net worth of assets was discovered. Since imputed income flows from assets were created through products of amounts and return rates, income flows could be positively, negatively or not at all biased.

Column 7 of Table 2 presents the average net asset flows by holding pattern for responders, and Column 8 exhibits the ratio of imputed to reported average flow. Half of the ratios meet or exceed 100 percent, and half are below it. However, the ratios fluctuate substantially, ranging 27 to 200 percent. In fourteen of

eighteen patterns, average imputed flows are within 40 percent of the reported averages. These patterns collectively account for 82 percent of all cases in the data. The fluctuation of imputed averages above and below those reported suggests no systematic bias in the imputation of asset flow. It may, however, suggest a large amount of imprecision (i.e., variability) associated with the imputation process.

Column 9 presents the ratio of total sample to reported average flows. Like Column 8, half the ratios meet or exceed 100 percent, and half are below it. Overall, average asset income flows are within 6 percent of the reported averages in ten of eighteen patterns. These cells account for 66 percent of all cases. In 6 of the remaining patterns, total averages are within 14 to 20 percent of those reported, and these cells account for 28 percent of all cases. Under this nonresponse model, there appears no systematic effect of imputed asset flow on the total data set.

Figure 2. Plot of Ave. Rates of Return in R vs Rates in NR



5.4 Income Flows from Individual Assets

We begin our investigation of asset income flows by measuring the contribution of imputed data to estimated total flow from individual assets. We decomposed the estimated total population flows for each asset into contributions of reporters and imputed cases. Imputed cases accounted for roughly 30 to 50 percent of estimated total population flows. The large contributions from imputed data suggest that population totals should not be estimated for asset income flow.

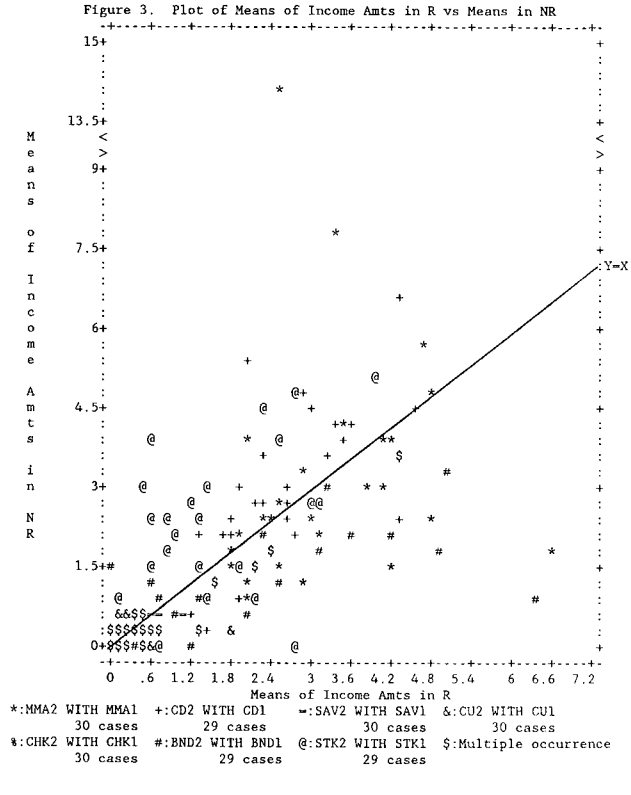
Next we examined the effect of imputation on average asset flows. Ratios of average imputed to reported flows varied substantially, depending on asset type. Average imputed flows exceed reported averages by 26 to 37 percent among CDs, checking accounts and stocks. On the other hand, mean imputed flows are 23 to 53 percent under reported means for savings accounts and bonds.

With respect to the total sample mean, total mean flows were within 14 percent of reported means for all assets except bonds. Bonds present a striking exception; here, the overall mean flow was 35 percent below that reported.

To check the plausibility of imputation, means of imputed flows were plotted against reported means, with each point representing a cell mean derived from our 30 cell nonresponse model. Figure 3 presents this plot, overlaid for all assets. Overall, the plotted points follow the Y=X line and are distributed about the line uniformly. However, specific assets behave differently. Plotted means of stocks (@) and CDs (+) congregate above the Y=X line, while those of bonds (#) generally fall below it. Thus, imputed stocks and CDs exhibit positive biases, imputed bonds display negative bias and the the other imputed assets exhibit no systematic bias.

Imputed asset income flows were products of imputed or reported asset amounts and imputed or reported return rates. Our analyses suggest that the positive bias for imputed asset amounts were offset by the negative bias of return rates for CDs and money market, checking and savings accounts. Imputed flows for these assets exhibited no systematic bias under our nonresponse model. The negatively biased return rates for imputed stocks and CDs did not completely offset the overstated asset amounts, producing positively biased flows. Finally, understated bond return rate imputations seem to overcompensate for overstated bond amount imputations, yielding negatively biased flows under our nonresponse model.

Figure 3. Plot of Means of Income Amts in R vs Means in NR



6. Summary and Conclusions

We have shown that imputation has modest to major effects on asset amount and flow in the NBS. There are two reasons for this. First, the extent of missing data, and hence imputation, was substantial. It ranged from 16 to 36 percent among asset amounts, and 22 to 65 percent among asset income flows. Also, the distributions of imputed values appear to be substantially different from reported distributions, in general. This affects estimates of population totals, means and aggregates (e.g., net worth from all assets).

To check the plausibility of imputation, two nonresponse models were employed. Assuming these models are true, small to modest positive biases were detected for net and individual asset amounts. Net return rates and net flows showed a slight to negligible negative bias. Results for individual asset return rates and flows were mixed. Slight to negligible negative biases were detected for individual return rates. For individual asset flows, bonds showed a slightly negative bias, stocks and CDs exhibited slight positive biases and other assets showed no systematic bias under our nonresponse model.

Analysts may choose not to analyze the NBS asset data because missing data and imputation for asset amounts and flows are substantial. However, the rates of item missing data in the NBS are not atypical for asset items in large sample surveys. If these data are analyzed, precautions should be taken.

We recommend that population totals for either amounts or flows of individual assets should not be estimated for survey data. Imputed cases contribute 30 to 50 percent of the total value to the usual Horvitz-Thompson estimates.

We also recommend extreme caution when estimating overall and subgroup mean asset amounts and income flows. Depending on the asset, biases could be slight to modest. One alternative would be to abandon the imputed values and employ a likelihood-based approach to parameter estimation with missing data (e.g., EM Algorithm). These and other approaches are described in Little and Rubin (1987).

Most importantly, we urge analysts of these data to develop and utilize explicitly their own nonresponse models. We have used two such models in this paper. Other models could be more appropriate for particular analysis. Since the true values of missing data can never be retrieved, assessment of imputation is ultimately subjective. In this sense, statistical inferences from these data are subjective as well.

Finally, we reemphasize the need for assessment of imputations in public use survey data sets. While such assessments can never be exhaustive, the results can stimulate healthy dialogue and attentiveness to issues concerning imputed data. In turn, this could lead to more appropriate analyses of data subject to imputation.

REFERENCES

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

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

David, M.H., Little R.J., Samuhel, M.E. and Triest, R.K. (1986) Alternative Methods for CPS Income Imputations. Journal of the American Statistical Association, Vol. 81, 29-41.

Czajka, J. (1984) Imputation Methodology for the New Beneficiary Survey, Mathematica Policy Research, Inc., Washington, D.C., 1984.

Little, R.J. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York.

Mattlin, J.A. (1987) Missing Data and Intermonthly Variation in Reports of Income and Assets in the New Beneficiary Survey: A Comparative Analysis. Institute for Survey Research, Temple University, 1987.

Santos, R.L. and Lazaro, C.G. (1987) Assessing Effects and Plausibility of Asset Imputation in the NBS. Institute for Survey Research, Temple University, 1987.

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

The 1982 New Beneficiary Survey: Users Manual. US Department of Health and Human Services, Social Security Administration, Office of Policy, Washington, D.C. Issued April 1986.