THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

ASSESSING THE EFFECT OF ALLOCATED DATA ON THE ESTIMATED VALUE OF TOTAL HOUSEHOLD INCOME IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP)

No. 244

Patricia J. Fisher
U.S. Census Bureau

U.S. Department of Commerce U.S. CENSUS BUREAU

Assessing the Effect of Allocated Data on the Estimated Value of Total Household Income in the Survey of Income and Program Participation (SIPP)

Patricia J. Fisher
U.S. Census Bureau

Assessing the Effect of Allocated Data on the Estimated Value of Total Household Income in the Survey of Income and Program Participation (SIPP)

In the Survey of Income and Program Participation (SIPP) item non-responses are imputed (or allocated), using a hot-deck procedure or, in some instances, are logically imputed using available information. These imputed values are then used in creating aggregate amounts, such as the value of total household income. In the SIPP, total household income is the sum of all income amounts reported or allocated by household members and includes wages and salaries earned, dividends, rental income, Social Security payments, income received from public assistance programs, Supplemental Security Income (SSI), and other sources. Any or all of these income sources could have been allocated and, to SIPP users, it is not well understood how these allocated amounts affect the created value of total household income.

Using the 2001 SIPP panel, this paper looks at the individual components of total household income and discusses the proportion allocated for each component.

Keywords: Imputation, Allocation, RHTOTINC, SIPP, data quality

1. Introduction

The Survey of Income and Program Participation (SIPP) survey is a longitudinal survey that produces national-level estimates for the U.S. resident population. The main objective of SIPP is to provide accurate and comprehensive information about the income and program participation of individuals and households in the United States, and about principal determinants of income and program participation.[1] The SIPP survey is often used for evaluating the effectiveness of government programs and the well being of the nation. SIPP collects household, family and person-level information on assets, health care, taxes, business and job income, participation in state and federal income-transfer programs and other demographics. The survey is designed to support longitudinal studies. This design feature allows for the analysis of selected dynamic characteristics of the population, such as changes in income, (Hisnanick and Walker, 2004), wealth (Gottschalck, 2006), participation of mothers in government programs, (Lugaila, 2005), disabilities (Steinmetz, 2006) and other characteristics.

The total monthly household income variable, (RHTOTINC), is widely used to impute other SIPP variables and therefore plays an important role in the data quality of SIPP. This analysis focuses on the allocated distribution of the components of the RHTOTINC variable that is calculated from summing all household members' incomes such as job, pension, social security, interest dividends and other income sources. It is not easy for users to determine how much of total household monthly income is allocated. This paper examines the allocation rate of each component of the total monthly household variable and their contribution to the total. This paper also examines the contribution of prior waves of allocation and the impact they might have on the overall allocation rates. And finally, this analysis discusses whether or not SIPP may be underestimating the hot deck allocation rate of the total monthly household income variable for waves 2 and higher as a result of using prior wave data.

The SIPP survey has many different users including public and private companies, academic researchers, other government agencies and policy makers. To keep the survey consistent over all users, allocations are included which are an integral part of the SIPP processing system. Almost every question asked in the SIPP survey has an unedited variable, an edited variable and an allocated variable. An allocated variable indicates whether or not the edited variable was imputed or allocated. An unedited variable is imputed if the response is invalid, such as, the

1 SIPP Users' Guide, Chapter 1

respondent answered, "don't know", refused to answer the question, the reported value is out of range, or the value is blank when it should be filled. In some cases, if one variable is imputed, the following variable will also be imputed to continue the pattern flow. For example, if a respondent answered "don't know" to holding a checking account and the edited variable was allocated to a 'yes', then the following question that ask "How much did you earn in interest?" is allocated, otherwise it would have been left blank since the respondent's interview never reached this question. Non-response is not limited to SIPP, other surveys such as the Current Population Survey (CPS) report that income data is also affected when respondents are unable to provide exact income information. [2]

During the 2001 panel, the nation underwent an economic transition. At the start of the panel, February 1, 2001, the Department of the Treasury reported the three month daily treasury yield was 5.87[3] percent. Towards the end of the panel, December 31, 2003, the three-month treasure yield was 0.95 percent.

1.1 SIPP Participants

The SIPP has a nationally representative sample of individuals 15 years of age and older in households in the civilian non-institutionalized population and these individuals are interviewed every 4 months (or wave) for 3 or 4 years for the entire panel. Each interview asks about the previous four months for its reference period. The 2001 SIPP panel consists of 9 waves beginning from February 2001 and ending January 2004.

1.2 Interviewing

Since the 1996 panel, computer aided interviewing (CAI) with laptop computers has been used rather than paper questionnaires, with the skip patterns built in. CAI allows use of prior wave data for editing missing data from later waves right in the instrument on laptop, thus lessening the need for subsequent longitudinal editing. However, editing and allocation also occur after SIPP interviews are completed. [4] The Census Bureau edits data for consistency, allocates missing data, and creates internal data files and public use files for each wave.

The SIPP survey allows respondents to report two jobs or businesses and up to ten out of 54 types of general income sources. If there were more than two jobs or more than ten sources of income, the survey asks the respondents to provide information on the largest source. This analysis recoded the job, business and the general income variables while keeping a tally of all the records that had at least one allocated value.

1.3 SIPP Allocation

Almost every variable in SIPP has an accompanying allocation variable also known as an allocation flag. Internal use data allocation flags has values of 0, 1, 2, 3, and 4. (See Box 1). Public use data allocation flags have values of 0, 1, 2, and 3. (See Box 2). The allocation flag of 4 in the internal use data converts to an allocation flag of 3 in the public use data.

Box 1: Internal Data

Allocation Flags

Value	Definition
0	Not in universe or no allocation
1	Allocated using the hot deck method
2	Cold deck allocation
3	Logical allocation using skip pattern logic
4	Carried over from prior wave because of respondent burden or current wave response is not valid. Not used on public use file.

An allocation value of '0' represents either (1) the edited variable is not in universe and thus allocation does not apply or (2) the edited variable is in universe and it has a valid response.

[2] D.Weinberg, "Income data quality issues in the annual social and economic supplement to the current population survey", 2004

[3] http://www.ustreas.gov/offices/domestic-finance/debt-management/interest-rate/yield_historical_2003.shtml

[4] SIPP User's Guide, Chapter 1

An allocation value of '1' indicates that the unedited variable did not have a valid response; the edited variable is allocated using a hot deck allocation procedure[5] and the allocated value is assigned to the edited variable. Since hot deck allocation is the most common of allocation, it is emphasized throughout this paper.

A value of '2' indicates that there was no data in the hot deck to use for allocation, thus a cold deck value was assigned to the edited variable.[6] Cold deck imputation is usually used rarely.

A value of '3' in the internal data file signifies that the edited variable was allocated logically. For example, if a respondent reported receiving joint checking interest, then logically they had a joint checking account. Logical imputation is rare in SIPP.

A value of '4' is shown only in the internal file and it identifies that the response was taken from the prior wave. This response could be stored from one of two ways (1) the respondent's recorded value in the prior wave or (2) the allocated value in the prior wave using hot deck, cold deck or logical imputation.

Box 2: Public Use File

Allocation Flags

Value	Definition
0	Not in Universe or no allocation
1	Allocated using the Hot Deck Method
2	Cold Deck Allocation
3	Logical Allocation: (1) Using skip pattern logic (2) Using prior wave data - (internal Flag of 4)

The public use file's allocation flags 0, 1 and 2 are similar to the internal data file's flags. However, the value of '3' in the public use file differs from the internal data file. For the public use file, a value of '3' can imply one of two things (1) the edited value was allocated logically in the current wave data or (2) the edited variable was allocated using prior wave data (the internal allocation flag of '4'). Here again, the prior wave data can be stored from either (1) the actual response from the respondent in the prior wave or (2) an allocated response stored in the prior wave using hot deck, cold deck or logical allocation based on the data at the time collected.

1.3 Methodology

The Census Bureau combines related questions and create recodes to make it easier for their users. These recodes are helpful since they cut down the amount of programming required as well as making it easier to understand each variable's characteristics. However, the recodes are not able to carry over allocation flags, thus no recodes were used in this analysis. To obtain allocation rates of the individual income sources that makes up the total monthly household income variable, it was necessary to use the Census Bureaus' internal files.

Each income source of the total monthly household income recode was evaluated using person weights. If an edited variable's allocation flag was 1, 2, 3 or 4 then the amount of the edited variable was tabulated accordingly. Wave 1, month 1 was tabulated using only allocated flags of 1, 2 and 3. Wave 5, month 1 was tabulated using flags of 1, 2, and 3 and then tabulated separately using flag 4.

The total household income variable formula is as follows:

Definition of Total Household Income=

Salary + General Income + Rental Property Income+ Job Income + Interest Income + Dividend Income+ Moonlighting Job + Business Income

The above formula consists of only variables from the core section of the SIPP; no topical module variables are used.

In most amount questions in the 2001 SIPP panel, a value of zero can represent either 'not in universe' or a valid zero amount. It was therefore necessary to obtain the correct universe by extracting out all the zeros that represent 'not in universe' and keeping only those that represent actual amounts The 2001 SIPP used computer assisted interviewing (CAI) which helped keep the skip patterns consistent in the SIPP interviews. The CAI interviewing does not allow interviewers to mistakenly skip questions like the paper interview allowed. Some questions are more

[5] See SIPP User's Guide 4
[6] See SIPP User's Guide 4

sensitive or more difficult to answer than others and thus are expected to have a higher missing value rate. If a respondent refused a question or didn't know the answer, a response from the same question either in the prior wave or month was allowed to fill the blank.[7] The hot deck allocation rate presented here would likely be underestimated if we only analyzed data using the allocation flags of 1, 2 and 3. If we examine the difference between waves for flags 1, 2 and 3 and separately look at the flag of 4 for wave 5, we can examine the difference by eye to determine if the any of the 4's in wave 5 are actually 1, 2 or 3s in earlier waves.

2.0 Using Prior Wave Data Allocation

In some cases it makes more sense to keep the prior wave data than to use the hot deck allocation. For example, income data from wages and salaries tend to be relatively stable over time compared to asset data such as interest and dividends, which can vary by quarter due to market fluctuations over time.

To reduce respondent burden, some questions are not re-asked in every wave. For example, the number of residential units in a house is asked only in the first interview and then each additional wave is updated with the same response and a '4', is assigned to the allocation variable for the internal data file. Intuitively, this flag of '4' indicates the household unit has not moved. In this circumstance using a flag of 4 reduces the need to re-ask the household how many units for the entire three-to-four year survey panel, thus helping the respondent as well as the interviewer.

The allocated value for an item in wave 1 may be carried through to future waves if the respondent does not provide a valid response, thus, allocating one response in one wave could impact the same response in future waves if the respondent does not provide a valid response sometime in between. Section 1.3 illustrated that the allocated flag value of '4' is not shown on the public use file. Users are unable to determine if the data with flag of '3's were allocated logically in the current wave or obtained from prior wave data, which may or may not have been allocated.

Two main effects that may result from carrying over prior wave data are first, whether or not the 4s are clearly visible, the analyst is likely to think of the 4s as data collected from respondents in the prior wave while in reality the data may have been allocated in the prior wave or even in several prior waves to the current wave which will then underestimate the overall understanding of hot deck imputation. Second, users of the public use files are not able to see flags with a value 4, and so are not able to perform a realistic analysis of hot deck allocation rates.

It is necessary to evaluate the prevalence of assigning previously hot deck imputed data to current missing data. Here, the first month of the first wave is compared to the first month of the fifth wave of the 2001. The first wave inherently does not have values of '4's for allocation flags; thus it is the only wave and month that can be evaluated without any prior data. The fifth wave was chosen because it was a year later and the respondents had sufficient time to experience the survey. There is likely an undesirable effect however, that the respondent also has acquired enough information in how the survey asks questions, especially sensitive questions to be able to avoid the skip patterns that lead them there.[8]

One benefit of using prior wave data instead of using a hot deck procedure is that the data are more consistent. For example, consider a respondent who does not report earnings in either the current or prior wave. The hot deck procedure in the prior wave might allocate the respondent a value of $1,000 of earnings a month, while the hot deck procedure in the current wave might allocate a different value, for example, of $5,000. Since hot decking imputation puts a donor at the bottom of the deck each time they are called, hot decking provides less consistency than using prior wave data allocation.

If the data was not imputed in the prior wave, then using the prior data provides even a greater benefit for data quality.

3.0 Results

A more complex analysis would require review each allocated variable from wave 1 and follow it up to wave 5 to determine if hot decking was used originally or somewhere in between. However that is beyond the scope of this paper. For this paper, comparisons are between two waves and not what went on between the waves.

3.1 Distribution of the Total Monthly Household Income Recode

Table 1 shows the distribution of each income source portion of the total household income variable,

[7] SIPP User's Guide, Chapter 4

[8] SIPP User's Guide, Chapter 1

RTOTINC, for the first month of wave 1 and wave 5 for the 2001 SIPP panel.

The income source that makes up the highest percent of the total monthly household income is salary from jobs, 69.1 percent for wave 1 and 70.4 percent for wave 5. The General Income source is the second highest percentage of total monthly household income with 14.6 percent in wave 1 and 16.5 percent in wave 5. The asset income sources, Interest Income, Property Income and Dividend Income make up 5 percent of the total monthly household income recode in wave 1 and 3 percent in wave 5.

Table 1. 2001 SIPP Distribution of Total Monthly Household Income by Individual Components for Wave 1 and Wave 5 for the First Month

Source of Income	Wave 1 Month 1	Wave 5 Month 1
Job	69.1%	70.4%
General Income	14.6%	16.5%
Business	10.9%	10.1%
Interest Income	2.2%	1.4%
Property Income	1.2%	0.9%
Dividend Income	1.6%	0.7%
Moonlighting	0.4%	0.1%
Total	100%	100%

3.2 Allocation of the Total Monthly Household Income Recode

Table 2 shows the share of persons for whom a given income source had an allocation flag of 1, 2 or 3 for the first month in wave 1 and for wave 5. The weighted proportions of each of the five allocation flags were calculated for wave 1 and wave 5.

As shown in Table 2, the majority of income sources in wave 5 were allocated less often than in wave 1. For example, allocation rates for job income falls significantly form 15.1 percent to 4.1 percent in wave 5. General income shows a significant drop from 21.8 percent in wave 1 to 10.4 percent for wave 5. The allocation rate for business income reduced 5 fold, from 21.0 percent in wave 1 to 4.7 percent in wave 5.

However, when prior wave data allocation is included in the assessment, then the allocation rates between waves vary less than shown in Table 2 and allocation rates are always higher in wave 5. Table 3 provides allocation rates by the share of persons with allocation flag 4 and with allocation flags 1, 2, 3, 4 for wave 5, month 1. Recall, the first month of wave 1 does not have prior data recorded, thus this month does not have allocation flags of 4.

Table 2. 2001 SIPP Percent of Persons with Allocation Flags of 1, 2, 3 [9]

Source of Income	Wave 1 Month 1 (col 1)	Wave 5 Month 1 (col 2)
Job	15.1%	4.1%+
General Income	21.8%	10.4%+
Business	21.0%	4.7%+
Interest Income	39.4%	35.9%
Property Income	22.9%	15.5%+
Dividend Income	33.5%	38.4%+
Moonlighting	18.7%	3.9%+
Total	26.7%	18.5%

As seen in Table 3, many of the allocation flags in Wave 5 have a value of '4', indicating that the data may have been carried over from the prior wave. We can examine the differences with an eye to determining how many 4's in Wave 5 are actually '1's, '2's, or '3's in earlier waves.

The data in table 3 suggest that wave 5's allocation based on flags 1, 2, and 3 (shown in table 2), under-report the hot deck, cold deck and logical allocation rates. For instance, the job income source in wave 5 had 18.7 percent of cases allocated using prior wave data, flag 4. Compared to table 2 only 4.1 percent of the job income source cases in wave 5 were allocated using a hot deck, cold deck, or logical imputation, compared to 15.1 percent in wave 1.

It is unclear how much of the 18.7 percent of the Job income source with a flag of 4 had originally been a flag of 1, but the answer lies somewhere in between 4.1 percent and 18.7 percent.

Note the values shown in column 1 for wave 5 in table 3 are more similar to the values shown in column 1 for wave 1 in table 2.

It is not possible to say without further research if the same cases in column 1 in table 2 are included in column 1 in table 3.

[9] + Indicates a signicant difference

Table 3. 2001 SIPP Percent of Persons with Allocation Flags of 4 and 1, 2, 3, 4 for Wave 5, Month 1[10]

Source of Income	Flag of 4 (col 1)	Flags 1, 2, 3, 4 (col 2)
Job	18.7%	22.8%
General Income	24.3%	34.7%
Business	25.2%	30.2%
Interest Income	20.4%	56.3%
Property Income	21.4%	36.9%
Dividend Income	21.6%	60.0%
Moonlighting	33.9.0%	37.8%
Total	24.2%	42.7%

Table 4 provides percentages of the total household monthly income by source of income.

The Job income source had 15.1% of the dollar amount allocated in wave 1 compared to 24.0 percent in wave 5. Compared to the 2000 Current Population Survey, 27.5 percent of earnings from the respondent's longest job was imputed.[11]

Interest income showed an anomaly, as the percent allocated with a flag of 4 in wave 5, 20.3% was less than the percent allocated in wave 5 with flags of 1, 2, or 3. Half of the interest amount was allocated in wave 1 (50.6%) versus two-thirds in wave 5 (62.2%).

The data in table 4 suggest that the original allocations (Flag values 1, 2 or 3) are still allocated in wave 5; that is, the bulk of those cases (or of the dollar amounts) are previously allocated with flags 1, 2 or 3. Many of these, in fact seem to have been allocated at the very beginning, in Wave 1, month 1.

The true hot deck allocation rate for wave 5 lies between two values, since the fraction of the cases (or dollars) that have a flag 4 which are really previously allocated flags 1,2, or 3 is not identifiable from these tables.

If allocation values of '4' are used over a lengthy time period, they may have implications regarding cross-sectional estimates as well as estimates of change between waves or years. High type-4 allocation rates would tend to bias measured changes toward zero since the prior wave data is used for the current wave. For example, if a region becomes depressed and there is a lot of job loss, the type '4' allocations would lead to an under-estimate of the effects, perhaps substantially. Type "4" allocation would tend to bias income amount estimates downward, especially if the allocated data were collected several months before, as there is no CPI adjustment for type '4' allocations.

4.0 Conclusion

The total household income variable (RHTOTINC) is one of the most used variables in predicting other variables when respondents are unable to provide a valid response. It may therefore have a large effect on the allocation of other variables. Knowing how much of the variable is actually data collected from respondents versus data allocated from other respondents based on similar characteristics, helps researchers understand the effect of allocation on the total household income variable as well as the other variables it effects. RHTOTINC is widely used in economic analysis and is one of the primary variables analyzed in the P-70 report series. Recognizing that a large fraction, 28.8%, of the total household monthly income is allocated, much of it carried over from previous waves, should inform the researcher with respect of characteristics of the error of their estimates.

There is more research to do. One possibility is to go back by wave until the very original response or allocation is found. Then the wave 5 data could be tabulated using the original allocation variables to measure the fraction from an actual response or from

[10] The column, "*Flag of 4*", is calculated by summing all the cases/persons with an allocation flag of 4 and then dividing it by the total number of cases that were in the universe for that income source. The column, *"Flags 1,2,3,4"* is calculated by summing up the total number of cases of the income source and dividing it by the total number of cases. This is the allocated rate for wave 5, month 1.

More than a fifth (23.5%) of the general income portion of the RHTOTINC recode was allocated in wave 1 with an allocation flag of 1, 2 or 3, compared to 10.8 percent in wave 5 for flags of 1, 2 and 3. But once flag 4 allocations are included, almost forty percent (39.2%) of the general amount income was allocated using prior wave data in wave 5.

Close to a quarter of people who reported property income (25.2%) had their value allocated in wave 1, compared to third (31.9%) in wave 5.

[11] Imputed income as percent of Total Income, (Barbara Atrostic, U.S. Census Bureau)

Table 4. Percent of total amount of monthly household income by flags 1, 2, 3 and 4 that were allocated by Source of Income[12]

	Wave 1, Month 1 Flag 1, 2, 3[13]	Wave 5, Month 1 Flag 1, 2, 3	Wave 5, Month 1 Flag 4	Total Wave 5 Flags 1, 2, 3, 4
Source of Income	% Total Amount Allocated	% Total Amount Allocated	% Total Amount Allocated	%Total Amount Allocated
Job	15.1%+	4.9%+	19.2%	24.0%
General Income	23.5%+	10.8%+	29.3%	39.2%
Business	28.2%+	5.6%+	33.5%	39.8%
Interest Income	50.6%+	41.9%+	20.3%	62.2%
Property Income	25.2%+	8.7%+	23.8%	31.9%
Dividend Income	32.7%+	15.4%+	21.3%	36.7%
Moonlighting	18.5%+	2.0%+	17.8%	19.8%
Total	18.8%	6.6%	22.4%	29.0%

the various forms of allocation. It would also be interesting to see how old the original response or allocation is in later waves of a panel. If data with allocation values of '4' have a large average 'age', that may have implications regarding cross-sectional estimates as well as estimates of change.

Some of the on going research on allocation involves replacing matrices with estimation models that rely less on hot decking methodology. Understanding and improving allocation methods may have a large effect on data quality by leading to better estimates of the variables in question. The SIPP has some extremely sensitive questions that many respondents find difficult difficult to answer. Knowing the allocation rate of the questions as well as the allocation of the variables that are estimating the question is a very important part of the data. Hopefully researchers will be encouraged to consider the effect of the allocations.

[12] The column, "*% Total amount allocated, flag 1,2,3*" is calculated by summing up the total of the component's income amount and dividing it by the total allocated sum of flags, 12, and 3 the component's income amount. The column, "*% Total amount allocated, flag 1,2,3,4*" is calculated by summing up the total of the component's income amount and dividing it by the total allocated sum for all flags. This is the allocated rate for wave 5, month 1.

[13] Significant difference between column 1 and 2

References

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

Atrostic, B., Kalenkoski, C., "Item Response Rates: One Indicator of How Well We Measure Income."

Freedman,V., Wolf, D., 1995., "A case study on the use of multiple allocation".

Gottschalck, A., 2004 "Dynamics of Economic Well-Being: Labor Force Turnover, 1996-1999".

Hisnanick, J., Walker, K, 2004. "Dynamics of Economic Well-Being: Movements in the U.S. Income Distribution, 1996-1999".

Lugaila,T., 2005. “Dynamics of Economic Well-Being: Participation of Mothers in Government Assistance Programs: 2001.

Steinmetz E., 2006. “Dynamics of Economic Well-Being: Americans with Disabilities: 2002”.

Weinberg D. 2004. "Income Data Quality Issues in the Annual Social and Economic Supplement to the Current Population Survey.”

Westat, Mathematica Policy Research, Inc, 2001. “Survey of Income and Program Participation User’s Guide.”

The.Department.of..Treasury, http://www.ustreas.gov/offices/domestic-finance/debt-management/interest-rate/yield_historical_main.shtml