THE EFFECTIVENESS OF OVERSAMPLING FOR LOW INCOME POPULATIONS IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION[1]

Tiwanda M. Allen, Rita J. Petroni, Rajendra P. Singh
Tiwanda M. Allen, U.S. Bureau of the Census, Washington, DC   20233

KEY WORDS: Auxiliary variables, analysis, sample size, variance

I. INTRODUCTION

A. Background

The goal of the Survey of Income and Program Participation (SIPP) is to provide policy makers with accurate and comprehensive information about the economic situation of persons and households in the noninstitutionalized U.S. population. Over the years, budget constraints dictated a reduction in the SIPP panel size. As data from the reduced panels became available, analysts found it more difficult to conduct meaningful analysis of government programs for subgroups such as the low income population. In response to analysts concerns about the diminished usefulness of the SIPP data to meet its goal, the Census Bureau pursued various budget initiatives to increase the sample to its original size and oversampling of minority populations. (King, 1990a.)

This paper describes the oversample design for the 1990 SIPP panel which the Census Bureau introduced in February through May 1990 and interviewed through June to September 1992. It examines the effectiveness of this oversample design in reaching the SIPP goal at both initial and later SIPP interviews. Results of this paper can provide guidance to organizations considering to oversample low income populations (e.g. households with income less than 125% of its poverty threshold).

B. Design of the 1990 SIPP Oversample Panel

The Census Bureau originally planned to introduce a 1990 SIPP panel of about 20,000 households selected with equal probability. Instead, the Bureau introduced a panel of 23,600 households which included an oversample of minority populations. Initially, we wanted to use income data to oversample the low income population. However, due to time constraints, this was not operationally feasible. As a result, the Census Bureau used demographic characteristics of household heads who were occupying the sample housing units during February to May 1989 as auxiliary variables, to select our oversample. These

characteristics are: Black (BLK), Hispanic (HIS), and female headed with no spouse present living with relatives (FHNSP). Such households tend to have higher low income rates than the general population. (King, 1990a.)

Table 1 displays the design of the 1990 SIPP oversample panel.

C. How Successful was the Oversample Design?

This paper examines the success of the Census Bureau's approach in increasing the number of low income cases and the impact of oversampling on the reliability of cross-sectional estimates at the beginning of the 1990 oversample panel. The oversample approach has been successful in increasing the number of low income cases at the beginning and end of a two and a half year panel. In addition, we found that addresses occupied by a BLK, HIS, or FHNSP head in February through May 1989 tended also to be occupied by a BLK, HIS, or FHNSP head in February 1990 through September 1992. The Census Bureau's oversample approach is also generally successful in increasing the reliability of low income-related and other SIPP estimates.

II. DATA ANALYSIS

A. Methodology

In the following sections we analyze the stability of the characteristics of addresses and housing units with respect to auxiliary variables and low income status and increases in the number of low income cases after one year and then after an additional two and a half years. We also analyze the reliability of various characteristics after one year. In the remainder of the paper we simply refer to "the characteristics of occupants of sample housing units with respect to auxiliary variables" as "auxiliary variables."

Section B presents evaluation of stability between Wave 1 of the 1989 panel and Wave 1 of the 1990 panel (Wave 1 is the interview months from February to May of 1989 and 1990, respectively); and between Wave 1 and Wave 8 of the 1990 panel (Wave 8 is the interview months from June to September of 1992). Section C examines increases in number of low income cases due to the 1989 panel cases at Waves 1

---

and 8 of the 1990 panel. Both sections examine unweighted counts.

Approximately 3500 addresses from the 1989 panel were initially interviewed in Wave 1 of the 1990 panel. However, as we match addresses between Wave 1 of the 1989 and Wave 1 of the 1990 panels, we loose addresses because not all of the necessary information is available for both. As a result, our analysis at different stages has fewer than 3500 cases.

In section D, we compare variances for the 1990 oversample design to variances when the 1989 panel cases are excluded. We computed the variances using the half sample replication option of VPLX. (Fay, 1990) For the 1990 oversample design, we used our normal SIPP weighting procedures. The weights include several adjustments to the baseweights. Two of these are an adjustment for combining samples from both the 1989 and 1990 panels and a raking ratio adjustment to account for population counts by age, race, sex, and household relationship. (King, 1990 b and c.) We derived weights for estimates which exclude the 1989 panel cases by:

• dividing out the combining and raking ratio adjustments for each original 1990 panel case,

• computing a new raking ratio adjustment, and

• multiplying by the new raking ratio adjustment.

B. Stability

Since the oversample cases from the 1989 panel were chosen based on our auxiliary variables, we first analyze the stability of these variables. The higher the proportion of cases falling into the same auxiliary variable category from one time to another, the higher the stability of the variable.

Table 2 shows that 89% of the addresses that were BLK-HIS-FHNSP in Wave 1 of the 1989 panel were also in the same group in Wave 1 of the 1990 panel (i.e., one year later). We also examined the stability of each of the variables separately. The BLK and HIS addresses are most stable with 94% and 87%, respectively, remaining in the same group. The "other" and FHNSP groups have about 70% of their addresses remaining in the same group through February to May 1990.

If the auxiliary variables are stable after one year, we would expect a similar stability after an additional two and a half years in sample. Since after the initial Wave 1 interview we follow Wave 1 persons instead of Wave 1 addresses.

Table 3 shows the stability of the variables from Wave 1 to Wave 8 of the 1990 panel. Between Waves 1 and 8, 92% of the addresses that were BLK-HIS-FHNSP remained in the same group.

Of the three characteristics chosen for our auxiliary variables, the one most likely to change from Wave 1 to Wave 8 is FHNSP. The BLK, HIS, and "other" groups had 99%, 98%, and 96%, respectively, remaining in the same group. The FHNSP has 74% in the same group. Our results are as expected. Changes in household membership as a result of events such as marriages, divorces, births, and deaths can account for changes in the auxiliary variables from Wave 1 to Wave 8.

From tables 2 and 3, we can calculate that over three and a half years about 82% of the households that were BLK-HIS-FHNSP headed in Wave 1 of 1989 were also BLK-HIS-FHNSP headed in Wave 8 of the 1990 panel. After three and a half years, only about 52% of such households are still classified as FHNSP.

In addition to the stability of the auxiliary variables we were also interested in the stability of income status with respect to these variables after a year and then after an additional two and a half years. We classify a household into low income status if the household income is less than 125% of its poverty threshold.

From table 4 we calculated that 71% of BLK-HIS-FHNSP headed households that had low income status in Wave 1 of 1989 had low income status a year later. For the "other" households with low income status at Wave 1 of the 1989 panel, 41% of the households maintained low income status a year later. As for the households above 125% of their poverty thresholds in Wave 1 of 1989, about 12% of both the BLK-HIS-FHNSP and "other" households had low income status in Wave 1 of the 1990 panel.

Similar low income analysis was done for the two and a half year period from Wave 1 to Wave 8 of the 1990 panel. From table 5, 70% of BLK-HIS-FHNSP headed households with low income status in Wave 1 of 1990 had the same status in Wave 8 of the 1990 panel. This is about the same percentage as the one year analysis. The results were also similar between the one year and two and a half year analysis for the BLK-HIS-FHNSP headed households with incomes above 125% of the poverty threshold in Wave 1 of the 1990 panel, but have low income status in Wave 8 of the 1990 panel. After two and a half years, about 55% of both the "other" households that had and did not have low income status in Wave 1 of the 1990 panel had low income status in Wave 8 of the panel.

Over three and a half years, about 50% of the BLK-HIS-FHNSP households with low income status in Wave 1 of 1989, had the same status in Wave 8 of the 1990 panel. Only 23% of the "other" households maintained low income status.

The above analyses display the success and stability of our auxiliary variables and income status when using the demographic characteristics of housing units. However, our original desire was to supplement our 1990 panel with only low income cases. Thus, we decided to simulate and analyze an oversample design based on income data. We then compared the stability of the income data to that of the current oversample design.

To perform this additional analysis, we used data from Waves 1 and 8 of the 1990 panel. After removing the supplemented 1989 panel cases, we determined the income status of the original 1990 panel households. The two and a half year analysis shows that 61% of the households with low income status in Wave 1 of 1990 had the same status in Wave 8 of the 1990 panel.

In developing a similar analysis for the supplemented 1989 panel cases, we found 67% of the households selected based on our auxiliary variables that had low income status in Wave 1 of 1990, had the same status in Wave 8. Actually, these cases have been in sample longer than two and a half years since they were originally interviewed in wave 1 of the 1989 panel.

Therefore, our current oversample design, based on the housing unit occupants' characteristics (BLK, HIS, FHNSP), provided better results than the originally planned design. This is explained by the fact that BLK, HIS, FHNSP populations have higher proportions of persons with low income than is found in the general population. BLK and HIS status of persons remains the same, and the change rate in FHNSP is low.

C. **Sample Size**

The 1989 panel cases increased the original 1990 panel size 17%. We calculated percentages of low income cases for both the cases taken from 1989 panel and the original 1990 panel. We found that for both sets of cases about 31% of BLK-HIS-FHNSP headed households are low income while only approximately 12% of the "other" households have low income status.

The BLK-HIS-FHNSP addresses from the 1989 panel are providing a 44% increase in the number of BLK-HIS-FHNSP low income households in Wave 1 of the 1990 panel while the "other" addresses are providing a 10% increase for its group. Totally, the 1989 panel addresses have increased the number of low income cases 26% for Wave 1 of the 1990 panel.

Similar results from the 1989 panel cases were obtained at Wave 8 of the 1990 panel.

D. **Reliability**

One goal of the oversampling was to reduce the variances of low income-related estimates without having a significant adverse affect on the variances of other SIPP estimates. Although our method increases the sample size for all population groups, the design introduces differential weights between cases from the original 1989 and 1990 panels. Since increased sample sizes decrease variances and differential weights increase them, we examine the actual variances to evaluate whether we met our goal.

We analyzed two sets of approximately 1700 cross-sectional national estimates and variances. One set was produced using the oversample panel cases, while the other set does not include the 1989 panel cases (the non-oversampling panel). The sets of estimates available are for the first quarter of 1990. This allows us to evaluate the reliability of estimates at the beginning of the 1990 panel.

Overall, variances for 74% of the 1700 estimates from the oversample design are smaller than the variances for non-oversample design. The majority (66%) of the variances from the oversample design are at least 10% smaller than the variances from the non-oversample design. The oversample approach has positively affected these estimates by decreasing their variances.

In addition to the above analysis, we compared the same variable characteristics for different populations such as the Total, Black, Hispanic, and persons aged 65 and over (65+), to see the affect the oversample approach is having on these different groups. Overall, we found that the variances for the oversample approach were smaller than for the non-oversample approach for the majority of the estimates analyzed for the Total, Black, Hispanic, and 65+ populations. These estimates included low income and non-low income type estimates. Therefore, in general, the oversample approach is improving the variances of low income estimates, without adversely affecting the general SIPP estimates.

Three examples of exceptional cases where the oversample design did not have a positive affect on variable characteristics of all the different populations are:

• Persons aged 16+ in the total population receiving cash benefits from means-tested programs.

• HIS persons aged 16+ receiving unemployment compensation.

• HIS persons and people in the total population participating in the food program for women, infants and children.

III. **RESULTS**

To analyze the effectiveness of the 1990 panel's oversample design we studied the stability of the auxiliary variables and income status relative to these

variables, the increase in sample size, and the reliability of our SIPP estimates.

The results from the analysis of the auxiliary variables we selected showed that the characteristics of the occupants of sample households or addresses are stable after one year and an additional two and a half years. After three and a half years in sample, 81% of the households that were BLK-HIS-FHNSP headed in Wave 1 of the 1989 panel were BLK-HIS-FHNSP headed in Wave 8 of the 1990 Panel. This stability is mainly due to the type of variables selected and the fact that for the two and a half year period SIPP follows Wave 1 persons rather than Wave 1 addresses.

The analysis of the stability of income status with respect to our auxiliary variables showed that, over three and a half years, 50% of the BLK-HIS-FHNSP headed households maintained low income status.

The 1989 panel addresses included in the 1990 panel increased the sample size about 17%. With this increase we were able to obtain approximately a 26% increase in the number of low income cases in both Wave 1 and Wave 8 of the 1990 Panel.

Generally the reliability of our 1990 first quarter low income type estimates have improved along with the reliability of our other 1990 first quarter SIPP estimates. Recall, 74% of the variances from the oversample design are smaller than the non-oversampling panel variances for 1700 cross-sectional estimates.

Initially we wanted to use income data to identify low income households. However, due to time and budget constraints we were unable. This paper showed that if we had used income data (as initially we wanted) instead of the selected auxiliary variables our results would have been less successful at later waves.

IV.   CONCLUSIONS

Results of the research presented here show that the SIPP 1990 oversampling method was successful for the SIPP cross-sectional estimation purposes both at the beginning and end of the panel.

The oversampling approach was successful in increasing the number of cross-sectional low income cases and improving the reliability of cross-sectional low income estimates without a significant adverse affect on other cross-sectional national estimates. To complete the analysis, we should:

* evaluate variances for estimates obtained later in the panel,
* research the impact on longitudinal variances, and
* research the impact of oversampling on cross-sectional and longitudinal variances when the

oversample design and non-oversample design are the same size.

These results also suggest that, at least when the goal is to oversample for low income households and follow them for two to three years or more, screening using income is not the best method. As long as low income rates remain stable for auxiliary variables, using auxiliary variables for which the characteristics of occupants of housing units are stable over time and correlated with low income, is the better approach. More specifically, if the low income rates are available and we know how many additional low income cases we need, we can estimate the size of our oversample. Since the low income rate is higher for the minority populations (BLK-HIS-FHNSP), using this group allows us to select a smaller sample to obtain more low income sample. Research is needed to determine which is the better method when the goal is to oversample low income households for a one time survey that is to be carried out close to the time of screening.

REFERENCES

Fay, R. (1990), "VPLX: Variance Estimates for Complex Samples," 1990 Proceedings of   the Survey Research Methods Section of the American Statistical Association.

King, K. (1990a), "SIPP: Restructuring the 1989 and 1990 Panels," Internal Census Bureau Memorandum for Documentation, January 31, 1990.

King, K. (1990b), "SIPP 90: Cross-Sectional Weighting Specifications for Wave 1," Internal Census Bureau Memorandum for Courtland from Waite, September 26, 1990.

King, K. (1990c), "SIPP 90: Cross-Sectional Weighting Specifications for Second and Subsequent Waves," Internal Census Bureau Memorandum for Courtland from Waite, November 5, 1990.

ACKNOWLEDGEMENTS

Table 1: The 1990 SIPP Oversample Panel.

Components of Oversample Panel	Number of Eligible Households
Households in addresses originally to be first interviewed in the 1990 panel.	19,700
Households associated with sample addresses which were to be first interviewed in February through May 1989 (i.e., households originally to be in the 1989 panel[2]) and were at that time headed by a BLK, HIS, or FHNSP.	2,700
Households in one-ninth of all other 1989[2] panel sample addresses.	1,200
Total	23,600

Table 2: Demographic Characteristics of Households Living at Sample Addresses in Wave 1 of the 1989 Panel Versus Wave 1 of the 1990 Panel.

1989 Panel, Wave 1	1990 Panel, Wave 1		
	Type of Address		
Type of Address	BLK-HIS-FHNSP	Other	Total
BLK-HIS-FHNSP	1458 89.17	177 10.83	1635
Other	418 29.33	1007 70.67	1425
Total	1876	1184	3060

Table 3: Demographic Characteristics of Households Living at Sample Addresses in Wave 1 of the 1990 Panel Versus Wave 8 of the 1990 Panel.

1990 Panel, Wave 1	1990 Panel, Wave 8		
	Type of Address		
Type of Address	BLK-HIS-FHNSP	Other	Total
BLK-HIS-FHNSP	1660 91.81	148 8.19	1808
Other	47 3.86	1172 96.14	1219
Total	1707	1320	3027

---

[2]   The Census Bureau attempted to interview households in all sample addresses from the 1989 panel in February 1989 through January 1990. After January 1990, we did not interview for the 1989 panel. However, for the 1990 oversample panel, we interviewed the 1989 panel households included in the 1990 oversample panel.

Table 4: Income Status of BLK-HIS-FHNSP Households in Wave 1 of the 1989 Panel Versus Wave 1 of the 1990 Panel.

| 1989 Panel, Wave 1 | | Income Level: 1990 Panel, Wave 1 | | | | Total |
| | | Below 125% | | Above 125% | | |
Income Level	Type of Address	BLK-HIS-FHNSP	Other	BLK-HIS-FHNSP	Other	
Below 125%	BLK-HIS-FHNSP	360	11	117	33	521
	Other	29	45	36	70	180
Above 125%	BLK-HIS-FHNSP	124	8	857	125	1114
	Other	74	82	279	810	1245
Total		587	146	1289	1038	3060

Table 5: Income Status of BLK-HIS-FHNSP Households in Wave 1 of the 1990 Panel Versus Wave 8 of the 1990 Panel.

| 1990 Panel, Wave 1 | | Income Level: 1990 Panel, Wave 8 | | | | Total |
| | | Below 125% | | Above 125% | | |
Income Level	Type of Address	BLK-HIS-FHNSP	Other	BLK-HIS-FHNSP	Other	
Below 125%	BLK-HIS-FHNSP	365	18	143	19	545
	Other	9	73	5	60	147
Above 125%	BLK-HIS-FHNSP	147	13	1005	98	1263
	Other	9	65	24	974	1072
Total		530	169	1177	1151	3027