Earned Income and Longitudinal Attrition in the SIPP and SPD

David Hall, Smanchai Sae-Ung, and Jan Tin, U.S. Census Bureau
Contact Author david.warren.hall@census.gov

Key Words: longitudinal survey, nonresponse, attrition bias, earned income, survey data quality, weighting, Social Security, administrative records

Abstract: Longitudinal surveys, like the Survey of Income and Program Participation (SIPP) and the Survey of Program Dynamics (SPD), experience growing attrition over the life of a survey panel. With each attriter, we lose information on variables such as earned income, which is commonly regarded as an important characteristic that is indicative of the socioeconomic well being of an individual and his/her family. In this study, we conduct a statistical evaluation of the longitudinal difference in earned incomes between the attriters and continuers in the SPD and the 1996 SIPP panel surveys using Social Security Administration (SSA) earnings records. We also analyze the difference between the earned income data collected from the SIPP and SPD surveys and those from SSA earnings records.

1. Introduction[1]

The Survey of Income and Program Participation (SIPP) is a longitudinal panel survey with the main objective of providing policy makers with comprehensive information about the income and program participation of the civilian noninstitutional population living in the United States. The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (commonly known as welfare reform) fundamentally altered the way government assistance was offered to low-income families. The Act mandated that the Census Bureau "continue to collect data on the 1992 and 1993 panels of the Survey of Income and Program Participation as necessary to obtain such information as will enable interested parties to evaluate the impact" of welfare reform. In response, the Census Bureau created the Survey of Program Dynamics (SPD) as a continuation of the 1992 and 1993 panels of the SIPP.

Both the SIPP and the SPD are nationally representative longitudinal surveys. Each SIPP panelist is followed and surveyed every four months over the life of the panel (typically three to four years). Each cycle of data collection is called a wave. The SPD survey picked up the sample from the 1992 and 1993 panels of SIPP for an additional six years of annual interviews (for a total of up to ten years of data). The longitudinal nature of the SIPP and SPD allows for measures of gross change, which is especially important when evaluating the impact of welfare reform. In contrast, a cross-sectional survey such as the American Community Survey (ACS) captures a snapshot in time and only allows for measures of net change.

One potentially significant issue with longitudinal surveys is the problem of attrition. As the panel ages, people cease to participate for a variety of reasons. Some may leave the panel through death or joining the armed services, others may simply lose interest or move away and cannot be relocated. As the percentage of nonresponse grows, so does the potential for bias in the survey results. As discussed in Kalton (1983), the population can be thought of as being divided into a stratum that would respond to a survey question and one that would not. The bias introduced by using the respondent sample mean to estimate the overall mean, if no compensation is made for nonresponse, is given by

$$B(\bar{y}_r) = E(\bar{y}_r) - \bar{Y} = \frac{M}{N}(\bar{Y}_r - \bar{Y}_m),$$

where $\bar{y}_r$ is the average value of a quantity of interest from the survey respondents (and $\bar{Y}_r$ is

[1] Disclaimer: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on issues are those of the authors and not necessarily those of the U.S. Census Bureau.

the corresponding population average), $\overline{Y}$ is the overall average value from everyone, M/N is the proportion of units in the population that do not respond (which can be thought of as the expected nonresponse rate), and $\overline{Y}_m$ is the (usually unobservable) average value from the nonrespondents in the population. This suggests that the average value from the survey respondents is approximately equal to the average value from the entire population when either (a) the nonresponse rate is small, or (b) the average value from the nonrespondents is close to the average value reported by the respondents. Unfortunately, the nonresponse rate for longitudinal surveys like the SIPP and SPD can be quite large, but the nonresponse bias may still be negligible if the respondents and nonrespondents are similar.

This paper extends the study begun by Vaughan and Scheuren (2002) on the differences of earned income distributions between SIPP and SPD respondents and nonrespondents. The values of earned income for both respondents and nonrespondents were obtained by matching to administrative records from the Social Security Administration (SSA). Vaughan and Scheuren evaluated the difference of annual median earned income of the 1993 SIPP panelists through the year 1999. In this paper, we extend our study to include the 1992 and 1996 panels of the SIPP as well as the SPD, and to cover the annual earned incomes through the year 2001. We also cover various percentiles of earned income (in addition to the median) to provide a statistical evaluation of the difference between the earned income data observed in the SIPP and SPD samples and those reported in the SSA administrative records. We use balanced repeated replication (BRR) per Fay (1989) for estimating the variances in our evaluation, and include various classifications based on race, ethnicity, sex, and age in the analysis.

2. Matching SIPP to SSA and Reweighting

The SIPP surveys requested social security numbers (SSNs) from respondents for research purposes. The Census Bureau was able to link SSNs from SIPP to the SSA Summary Earnings Records (SER) for adults age fifteen and older for about 89% of SIPP 1992 wave 1 respondents, 88% of SIPP 1993 wave 1 respondents, and 85% of SIPP 1996 wave 1 respondents. These matched records allowed us to use both SSA information, such as earned income, and SIPP and SPD data for a combined analysis.

By only using respondents to whom we could match SSNs, however, we could be introducing another bias into the analysis. Are the people who report SSNs different from the people who do not provide SSNs? To avoid this concern, we reweighted the matched cases to raise them back up to the original SIPP weighted control levels. Similar to what was done previously, we raked the SIPP linked cases to the same seven dimensions used by Vaughan and Scheuren: interview type (2 levels), age (2 levels), race and ethnicity (3 levels), household relationship (3 levels), citizenship (3 levels), family welfare ratio (2 levels), and total family monthly income (2 levels). This raking was done separately for each of the three SIPP panels 1992, 1993, and 1996.

We compared the original SIPP samples with the reweighted cases to assess the adjustment. We looked at education level, family size, family welfare recipiency, employment status, personal earnings, and family income. These variables were not specifically accounted for in the raking process, but the agreement at the national level between the original and the matched cases was generally very good.

We also reweighted the three sets of replicate weights, one for each SIPP survey. We will use the replicate weights for variance estimation later when we make comparisons.

3. Attriters versus Continuers

Once we reweighted the SIPP respondents to account for valid SSN matches, we could begin analyzing the differences between survey attriters and continuers. We defined an *attriter* as an original sample person who was not interviewed during the final wave of data collection for the panel. These people were interviewed in wave 1 but had dropped out of the survey at some point during the panel and never reentered. A *continuer*, in contrast, was someone who was interviewed in both the first wave and the last wave of the survey (but may or may not have been interviewed during all of the waves in between). We excluded cases in which the person left the survey universe through death, institutionalization, emigration, or by moving into an armed forces barracks, as well as people who joined the survey after the first wave of interviewing (for example, by marriage or birth).

By this definition of attrition, about 18% of wave 1 respondents age fifteen and older with

valid SSN matches in the SIPP 1992 panel were attriters. That value was about the same (20%) for the SIPP 1993 panel, but jumped up to 32% for the SIPP 1996 panel. That escalation may be partly due to the increase in panel length (12 waves of interviews vs. 9 or 10 for the earlier panels) or a general decrease in the public willingness to participate in government surveys [Atrostic et al. (1999)]. Thus it may be surprising at first to see that the attrition rate for SPD was only 33% given that this survey lasted for as many as ten years (1992-2002). However, the Census Bureau made great efforts to keep the attrition rate as low as possible by offering incentives for participation and going back to households that had dropped out of the survey during the earlier SIPP 1992 and 1993 panels and persuading them to return and participate in the SPD.

The attrition rates among blacks and Hispanics are substantially higher than the overall rate. The black attrition rates are 27%, 30%, 39%, and 40% for the SIPP panels 1992, 1993, 1996, and the SPD, respectively. Similarly, the attrition rates for Hispanics are 25%, 28%, 37%, and 39%. Overall, the attrition rates for males are slightly higher than for females, while the attrition rates among adults age 25 and younger are considerably higher than those for older adults.

It is also interesting to note that the overall attrition rates for the people to whom we could not match SSNs was 10% to 15% higher than for those to whom we could produce a valid SSN match for each of the four surveys.

4. Income Comparisons Between SIPP/SPD and SSA

Before we proceed to compare the annual earned incomes of continuers and attriters, we thought it would be prudent to check how well the SSA incomes reflected those reported in the SIPP and SPD. We could only make these comparisons with continuers (since, by definition, we do not have all of this information for attriters). However, we restricted our comparisons even further. A sizeable proportion of persons reported zero earned income (which we defined to consist of wages and salary as well as income from self-employment and farming). At the other extreme, large values of annual earned income are topcoded for disclosure reasons to protect confidentiality. We limited our comparisons to continuers age eighteen and older who were interviewed in every wave of data collection (self, proxy, or imputed) and who had a positive earned income value less than the top-code. This subgroup of continuers encompasses low- and middle-income people, who generally are more important for evaluating socioeconomic well being than high-income people.

We looked specifically at the annual earned incomes for years 1996-1999 from the 1996 SIPP panel and the years 1992-2001 from the SPD. We estimated the three quartiles of the income distributions from both the survey and the SER records. The quartile estimates were usually close, but often statistically significantly different. In particular, the percent differences for the medians were typically within 10%, as were those of the 75^{th} percentiles. The percent differences for the 25^{th} percentiles were generally larger (up to 15% or more) and usually significant. Table 1 displays a typical income distribution comparison.

We were encouraged to find rather strong correlations between the annual earned incomes as reported in the SSA data and the SIPP/SPD data. This held true when limiting our analysis to subgroups of blacks, Hispanics, males, and females as well (with correlations ranging from 0.67 to 0.87). However, there were numerous cases in which a respondent reported relatively little earned income to the SIPP or SPD yet had a large value of earned income in the SER or vice versa. These discrepancies could stem from a variety of sources, such as differences in how self-employment is defined, misreporting, the limited number of jobs reportable to the SIPP and SPD (up to four), and the use of imputation for item nonresponse in the SIPP and SPD.

If you assume the true value of annual earned income is obtained from the SSA administrative records, then the difference between the SIPP/SPD earned income and the SSA earned income can be thought of as measurement error. Our results indicate that measurement error is present in the SIPP and SPD and may affect data interpretation. Future research will attempt to model this relationship between the measurement error and earned income using key demographic and socio-economic variables. This should enable us to reduce the bias in the SIPP and SPD annual earned income data with respect to the assumed true value.

5. Income Comparisons Between Continuers and Attriters

Once we were fairly satisfied with the earned income data from the SSA as a surrogate for the SIPP and SPD earned incomes, we continued with our comparisons of continuers and attriters. We compared annual earned incomes from the SER for the years 1992 through 2001 for all adults age eighteen and over for which we had a valid matching SSN in the SPD and the SIPP 1992, 1993, and 1996 panels. We also broke the analysis down by race, ethnicity, sex, and age groups. We converted all income amounts to 2001 dollar values to account for inflation and limited our analysis to those with positive annual earned income. We tested comparisons of various percentiles of the continuer's and attriter's annual earned income distributions (ranging from the 5th to the 99th) at a 10% level of significance.

The shapes of the frequency distributions of annual earned income among all continuers and attriters for the SPD and the SIPP 1992, 1993, and 1996 panels are all similarly skewed to the right as typified by Figure 1. The frequency spikes at the far right end are due to the top-coding of the annual earned incomes.

The median annual earned income for all attriters was 10% to 25% lower than the corresponding median annual earned income for all continuers between the years 1992 and 2001 for the 1992, 1993, and 1996 SIPP panels. While the gap tended to close during that time period (see Figure 2 for a typical example), all of the differences were statistically significant for all years. The median annual earned income for all SPD attriters, on the other hand, ranged from 0% to 10% lower than the corresponding median annual earned income for all SPD continuers, with the final three years (1999-2001) showing no significant difference between the two groups. This distinction is most likely due to the nature of how the SPD sample was created and maintained. As mentioned earlier, some households that had dropped out of the original SIPP 1992 and 1993 panels were brought back to the SPD sample. But the vast majority (approximately 84%) of the SPD sample were continuers from the 1992 and 1993 SIPP panels.

Similar patterns appeared when we looked at subgroups of the survey populations as well. Continuers tended to have higher median annual earned incomes than the corresponding set of attriters in the three SIPP panels, although the differences were not always significant (see Table 2 for an example). Older persons tended to be the exception to this rule in that attriters often had higher median annual earned income than did continuers, though not usually significantly greater due to the small sample size. Males tended to show a greater difference between continuer's and attriter's median annual earned incomes than did females. The differences in median annual earned income between continuers and attriters within SPD were generally not significant for most of these domains (again, probably because the majority of attriters in the SPD were classified as continuers in the SIPP 1992 and 1993 panels).

The differences in median annual earned income between continuers and attriters tended to decrease beginning in 1996 in the three SIPP panels. This pattern generally did not extend to percentiles far above and below the median, however.

6. Discussion and Conclusion

Earned income is commonly regarded as an important characteristic indicative of the socioeconomic well being of an individual and his/her family. The attrition rate in a longitudinal panel survey like the SIPP and SPD is often high, so that the potential bias induced by the difference between continuers and attriters should not be ignored. In this study, we evaluated the difference between SIPP and SPD continuers' and attriters' annual earned income as reported in the SSA administrative records. Our evaluation covered the SPD and the 1992, 1993, and 1996 panels of the SIPP over the years 1992 through 2001. We considered subgroups of the population covering race, ethnicity, sex, and various age groups. We based our evaluation on direct estimates from the observed survey data and the data reported on the SER.

By assuming the annual earned income as reported to the SSA to be the true value, our statistical analysis indicates that the measurement error in the SIPP and SPD is significant and should be dealt with in terms of biases it may introduce. This SIPP/SPD measurement error appears to be correlated with the annual earned income reported to the SSA. This is contrary to the assumption of statistical independence between measurement error and the true value commonly made in classical modeling [Fuller (1987)].

We evaluated the difference between various percentiles of annual earned incomes of the SIPP and SPD continuers and attriters. Their differences, as reported in the SER, are more often than not statistically significant across the time span between 1992 and 2001 for the 1992,

1993, and 1996 SIPP panels. This suggests that it may be useful to use the SSA administrative records to improve the SIPP and SPD nonresponse weight adjustment and missing data imputation process.

Based on the above discussion, we plan to construct a model of the relationship between measurement error and annual earned income. We will attempt to derive a statistical relationship using SSA administrative records and other key demographic and socioeconomic variables to estimate the systematic and random components of the measurement error. An adequate model could be used to reduce the bias in the SIPP and SPD annual earned income data. We also plan to derive a relationship between annual earned income of continuers and attriters using certain key demographic and socio-economic variables to estimate the systematic and random components of this difference. We could use the statistical model to improve our SIPP and SPD nonresponse weight adjustment and imputation processes with respect to annual earned income.

7. Acknowledgement

We would like to gratefully thank Denny Vaughan and Fritz Scheuren for their invaluable guidance in the course of this research.

8. References

Atrostic, B. K., Bates, N., Burt, G., Silberstein, A., and Winters, F. (1999). *Nonresponse in Federal Household Surveys: New Measures and New Insights*. International Conference on Survey Nonresponse, Portland, Oregon.

Deming, W. E., and Stephan, F. F. (1940). *On a Least Squares Adjustment of a Sample Frequency Table When the Expected Marginal Totals are Known*. Annals of Mathematical Statistics.

Fay, R. E. (1989). *Theory and Application of Replicate Weighting for Variance Calculations.* Proceedings of the Survey Research Methods Section. American Statistical Association, 212-217.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley and Sons, Inc., New York.

Kalton, G (1983). *Compensating for Missing Survey Data*. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

Vaughan, D., and Scheuren, F. (2002). *Longitudinal Attrition in the SIPP and SPD.* Proceedings of the Survey Research Methods Section. American Statistical Association, 3559-3564.

U. S. Census Bureau (2001). *Survey of Income and Program Participation Users' Guide.* Washington, DC.

U. S. Census Bureau (2004). *Survey of Program Dynamics Users' Guide (Revised).* Washington, DC.

U. S. Congress (1996). *Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PL No. 104-193).*

Table 1: Comparison of annual earned incomes between the 1996 SIPP panel and the SER for years 1996 and 1999

Adult Subgroup	25th Percentile in 1996		50th Percentile in 1996		75th Percentile in 1996		SIPP-SSA Correlation
	Difference	% Difference	Difference	% Difference	Difference	% Difference	
Adults 18+	$1400*	13.0	$650*	3.0	-$410*	-1.1	0.83
Blacks 18+	$760*	8.4	$440	2.4	-$1000*	-3.2	0.78
Hispanics 18+	$930*	10.0	-$19	-0.1	-$830	-3.0	0.81
Females 18+	$880*	11.0	$30	0.2	-$220	-0.7	0.87
Males 18+	$2100*	16.0	$720*	2.7	-$220	-0.5	0.78
Adult Subgroup	25th Percentile in 1999		50th Percentile in 1999		75th Percentile in 1999		SIPP-SSA Correlation
	Difference	% Difference	Difference	% Difference	Difference	% Difference	
Adults 18+	$1400*	11.0	-$250*	-1.0	-$1800*	-4.6	0.79
Blacks 18+	$1700*	15.0	-$130	-0.6	$300	0.9	0.69
Hispanics 18+	$1600*	14.0	-$450	-2.2	-$1500*	-4.5	0.73
Females 18+	$1100*	11.0	$6	0.0	-$910*	-2.7	0.82
Males 18+	$1600*	9.5	-$730*	-2.4	-$1900*	-4.1	0.73

* - difference is significant at the 10% level

Figure 1: Distribution of annual earned income for SPD *attriters* age 18 and over for years 1993, 1996, 1999, and 2001

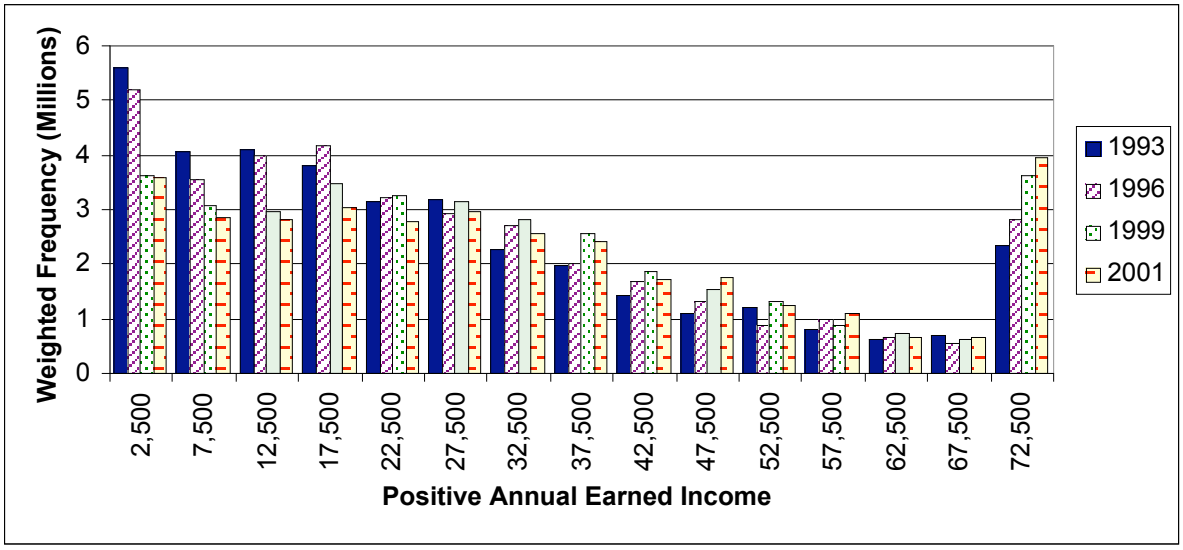

Figure 2: Median annual earned income (SSA data) for all SIPP 1993 continuers and attriters, age 18 and over

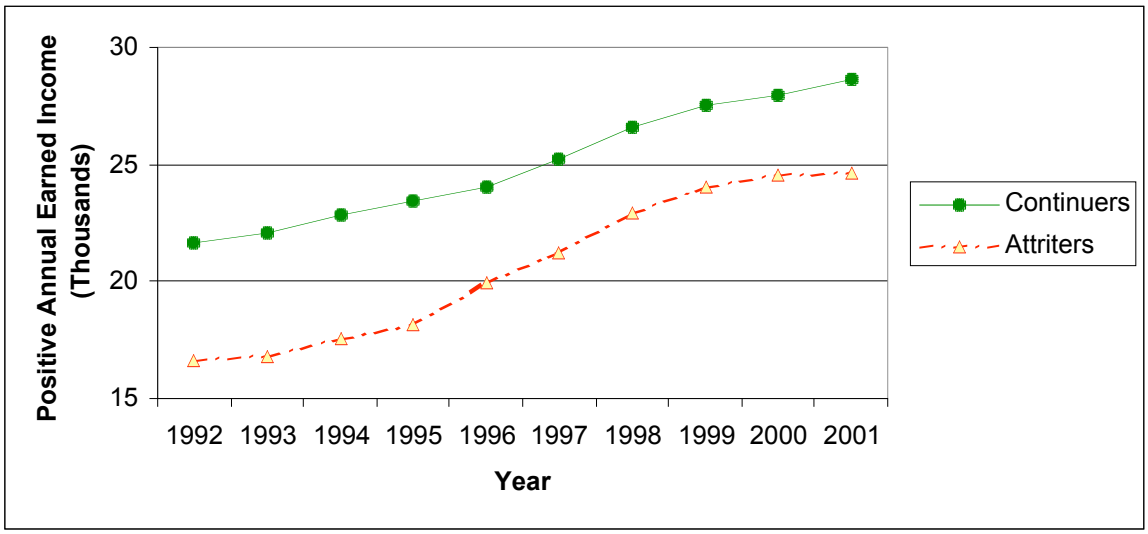

Table 2: Median annual earned income (SSA data) for SIPP 1992 continuers and attriters for years 1992 and 2001

Adult Subgroup	1992 (in 2001 Dollars)			2001		
	Continuers	Attriters	% Difference	Continuers	Attriters	% Difference
All Adults 18+	$22,000	$17,000	-25*	$29,000	$25,000	-13*
Blacks 18+	$18,000	$14,000	-19*	$22,000	$21,000	-8
Hispanics 18+	$17,000	$14,000	-20*	$23,000	$20,000	-15*
Females 18+	$17,000	$14,000	-18*	$23,000	$21,000	-8*
Males 18+	$29,000	$19,000	-35*	$36,000	$29,000	-19*
Adults 18-25	$11,000	$9,600	-11*	$27,000	$24,000	-9*
Adults 26-30	$23,000	$19,000	-19*	$30,000	$26,000	-13*
Adults 31-45	$29,000	$22,000	-24*	$33,000	$26,000	-21*
Adults 46-65	$25,000	$24,000	-4	$21,000	$22,000	7
Adults 66+	$5,800	$6,800	16	$4,400	$4,900	12

* - difference is significant at the 10% level